【書類名】　特許願
【整理番号】　0290541104
【提出日】　平成14年11月13日
【あて先】　特許庁長官殿
【国際特許分類】　G06F 17/30
【発明者】
　【住所又は居所】　東京都品川区北品川６丁目７番３５号　ソニー株式会社内
　【氏名】　大野　潮満
【特許出願人】
　【識別番号】　000002185
　【氏名又は名称】　ソニー株式会社
【代理人】
　【識別番号】　100082131
　【弁理士】
　【氏名又は名称】　稲本　義雄
　【電話番号】　03-3369-6479
【手数料の表示】
　【予納台帳番号】　032089
　【納付金額】　21,000円
【提出物件の目録】
　【物件名】　明細書　1
　【物件名】　図面　1
　【物件名】　要約書　1
　【包括委任状番号】　9708842
【プルーフの要否】　要

SONY 1311
US

日本国特許庁
JAPAN PATENT OFFICE

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office.

出願年月日
Date of Application: ２００２年１１月１３日

出願番号
Application Number: 特願２００２－３２９４９２

[ST. 10/C]: [ＪＰ２００２－３２９４９２]

出願人
Applicant(s): ソニー株式会社

２００３年 ８月１１日

特許庁長官
Commissioner,
Japan Patent Office 今井康夫

出証番号 出証特２００３－３０６４３７４

【書類名】　明細書

【発明の名称】　情報処理装置および方法、記録媒体、並びにプログラム

【特許請求の範囲】

【請求項１】　サイトを構成するページのデータを取得する取得手段と、

前記取得手段により取得された前記ページのデータを用いて、前記ページ内に出現する単語を抽出する抽出手段と、

前記抽出手段により抽出された前記単語が、前記ページ内で出現する回数をカウントするカウント手段と、

前記取得手段で取得されたページ間のリンク構造を解析し、前記カウント手段によるカウントの値を用いて、リンク関係にある前記ページ間の第１の重みを生成する第１の生成手段と、

前記第１の生成手段により生成された前記第１の重みを用いて、所定のページとリンク関係にある他のページ同士のページ間の第２の重みを生成する第２の生成手段と、

前記第２の生成手段により生成された前記第２の重みを用いて、ＳＤＦデータまたはＣＤＦデータの少なくとも一方のデータを生成する第３の生成手段と、

前記第３の生成手段により生成された前記データを用いて、ＩＳＤＦまたはＩＣＤＦの少なくとも一方に基づくページモデル拡張処理により所定の値を算出する算出手段と

を含むことを特徴とする情報処理装置。

【請求項２】　前記算出手段により算出された前記所定の値を用いて、前記取得されたページ間の所定のページ間の関連度を算出する第２の算出手段を

さらに含むことを特徴とする請求項１に記載の情報処理装置。

【請求項３】　前記第２の生成手段が、前記所定のページをリンク元とし、そのリンク元からリンクが張られているリンク先のページ同士のページ間の前記第２の重みを算出する場合、前記第３の生成手段は、前記ＳＤＦデータを生成し、前記算出手段は、前記ＩＳＤＦに基づくページモデル拡張処理により前記所定の値を算出し、

前記第２の生成手段が、前記所定のページをリンク先とし、そのリンク先にリ

ンクを張っているリンク元のページ同士のページ間の前記第２の重みを算出する場合、前記第３の生成手段は、前記ＣＤＦデータを生成し、前記算出手段は、前記ＩＣＤＦに基づくページモデル拡張処理により前記所定の値を算出し、

前記第２の生成手段が、前記所定のページをリンク元とし、そのリンク元からリンクが張られているリンク先のページ同士のページ間の前記第２の重みと、前記所定のページをリンク先とし、そのリンク先にリンクを張っているリンク元のページ同士のページ間の前記第２の重みをそれぞれ算出する場合、前記第３の生成手段は、前記ＳＤＦデータと前記ＣＤＦデータをそれぞれ生成し、前記算出手段は、前記ＩＳＤＦと前記ＩＣＤＦに基づくページモデル拡張処理により前記所定の値を算出する

ことを特徴とする請求項１に記載の情報処理装置。

【請求項４】　前記算出手段は、前記所定のページ内における所定の単語の出現回数と、前記所定のページと前記第２の生成手段で生成されたリンク関係にあるページのうち、前記所定の単語を含むページに対応する前記第３の生成手段により生成された前記データを用いた演算により、前記所定の値を算出する

ことを特徴とする請求項１に記載の情報処理装置。

【請求項５】　前記第２の算出手段により算出された前記関連度を記憶する記憶手段と、

所定のページに関連があるページの情報の提供が要求された場合、前記記憶手段に記憶されている前記関連度を参照して、前記所定のページと関連度が高いページの情報を提供する提供手段と

をさらに含むことを特徴とする請求項１に記載の情報処理装置。

【請求項６】　前記提供手段は、前記情報を提供する際、前記所定のページと関連する広告に関する情報も提供する

ことを特徴とする請求項５に記載の情報処理装置。

【請求項７】　サイトを構成するページのデータを取得する取得ステップと、

前記取得ステップの処理で取得された前記ページのデータを用いて、前記ページ内に出現する単語を抽出する抽出ステップと、

前記抽出ステップの処理で抽出された前記単語が、前記ページ内で出現する回数をカウントするカウントステップと、

前記取得ステップの処理で取得されたページ間のリンク構造を解析し、前記カウントステップの処理によるカウントの値を用いて、リンク関係にある前記ページ間の第１の重みを生成する第１の生成ステップと、

前記第１の生成ステップの処理で生成された前記第１の重みを用いて、所定のページとリンク関係にある他のページ同士のページ間の第２の重みを生成する第２の生成ステップと、

前記第２の生成ステップの処理で生成された前記第２の重みを用いて、ＳＤＦデータまたはＣＤＦデータの少なくとも一方のデータを生成する第３の生成ステップと、

前記第３の生成ステップの処理で生成された前記データを用いて、ＩＳＤＦまたはＩＣＤＦの少なくとも一方に基づくページモデル拡張処理により所定の値を算出する第１の算出ステップと、

前記第１の算出ステップの処理で算出された前記所定の値を用いて、取得されたページ内の所定のページ間の関連度を算出する第２の算出ステップと

を含むことを特徴とする情報処理方法。

【請求項８】　サイトを構成するページのデータを取得する取得ステップと、

前記取得ステップの処理で取得された前記ページのデータを用いて、前記ページ内に出現する単語を抽出する抽出ステップと、

前記抽出ステップの処理で抽出された前記単語が、前記ページ内で出現する回数をカウントするカウントステップと、

前記取得ステップの処理で取得されたページ間のリンク構造を解析し、前記カウントステップの処理によるカウントの値を用いて、リンク関係にある前記ページ間の第１の重みを生成する第１の生成ステップと、

前記第１の生成ステップの処理で生成された前記第１の重みを用いて、所定のページとリンク関係にある他のページ同士のページ間の第２の重みを生成する第２の生成ステップと、

前記第２の生成ステップの処理で生成された前記第２の重みを用いて、ＳＤＦデータまたはＣＤＦデータの少なくとも一方のデータを生成する第３の生成ステップと、

前記第３の生成ステップの処理で生成された前記データを用いて、ＩＳＤＦまたはＩＣＤＦの少なくとも一方に基づくページモデル拡張処理により所定の値を算出する第１の算出ステップと、

前記第１の算出ステップの処理で算出された前記所定の値を用いて、取得されたページ内の所定のページ間の関連度を算出する第２の算出ステップと

を含むことを特徴とするコンピュータが読み取り可能なプログラムが記録されている記録媒体。

【請求項９】　サイトを構成するページのデータを取得する取得ステップと、

前記取得ステップの処理で取得された前記ページのデータを用いて、前記ページ内に出現する単語を抽出する抽出ステップと、

前記抽出ステップの処理で抽出された前記単語が、前記ページ内で出現する回数をカウントするカウントステップと、

前記取得ステップの処理で取得されたページ間のリンク構造を解析し、前記カウントステップの処理によるカウントの値を用いて、リンク関係にある前記ページ間の第１の重みを生成する第１の生成ステップと、

前記第１の生成ステップの処理で生成された前記第１の重みを用いて、所定のページとリンク関係にある他のページ同士のページ間の第２の重みを生成する第２の生成ステップと、

前記第２の生成ステップの処理で生成された前記第２の重みを用いて、ＳＤＦデータまたはＣＤＦデータの少なくとも一方のデータを生成する第３の生成ステップと、

前記第３の生成ステップの処理で生成された前記データを用いて、ＩＳＤＦまたはＩＣＤＦの少なくとも一方に基づくページモデル拡張処理により所定の値を算出する第１の算出ステップと、

前記第１の算出ステップの処理で算出された前記所定の値を用いて、取得され

たページ内の所定のページ間の関連度を算出する第２の算出ステップと

をコンピュータに実行させることを特徴とするプログラム。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】

本発明は情報処理装置および方法、記録媒体、並びにプログラムに関し、特に、ネットワーク上で開設されているホームページなどの検索に用いて好適な情報処理装置および方法、記録媒体、並びにプログラムに関する。

【０００２】

【従来の技術】

近年、インターネットの普及により、そのインターネット上で開設されているホームページの数も増大しつつある。それらのホームページは、企業だけでなく、個人ユーザも開設しているため、その数は、膨大なものとなっている。それら膨大な数のホームページから、ユーザが所望の情報を掲載したホームページを探し出すということは大変な手間がかかることであった。

【０００３】

そのような手間を省くために、キーワードなどを入力するだけで、所望のホームページが検索できるような、俗に検索エンジンなどと称されるホームページ、例えば、Ｙａｈｏｏ（商標）、ｇｏｏ（商標）、Ｅｘｃｉｔｅ（商標）、Ｇｏｏｇｌｅ（商標）、Ｎｅｔｓｃａｐｅ（商標）がサービスの提供を開始している。

【０００４】

これらの検索エンジンは、ユーザが入力したキーワードを含み、キーワードの特徴に近い類似したホームページを探す際に適しているが、その検索結果以外にもユーザが所望するページが多い。

【０００５】

そのため、幾つかの検索エンジンでは、関連ページ検索などと称される関連ページ検索エンジンのサービスを開始している。例えば、特許文献１や、Ｇｏｏｇｌｅの検索結果の各々ページに対する関連ページ検索、Ｇｏｏｇｌｅ　Ｔｏｏｌｂａｒの関連ページ検索ボタン、Ｎｅｔｓｃａｐｅ　Ｎａｖｉｇａｔｏｒなどの

ブラウザに表示される関連サイト検索ボタンなどがある。

【０００６】

【特許文献１】

特開２００２－１４９６９８号公報（第４―７頁）

【０００７】

【発明が解決しようとする課題】

関連ページ検索エンジンを用いた検索は、ユーザが閲覧中のページ、あるいは検索エンジンの検索結果の所定のページに対して関連するページが検索される。その検索は、WWW（World Wide Web）のリンク構造を考慮するものもあったが、関連ページの検索が必ずしも精度良く行われているとは限らなかった。

【０００８】

これは、従来のページの特徴抽出によるページモデルの生成は、関連ページ検索ではなく、検索エンジン、つまり、入力されるキーワードや自然言語と検索対象となるページとの類似度を求めるための手段であったため、関連ページを検索する場合のページの特徴抽出には適していなかったためである。関連ページ検索では関連ページ検索に適したページの特徴抽出に基づくページモデルの生成が必要である。

【０００９】

本発明はこのような状況に鑑みてなされたものであり、リンク構造のうち、兄弟関係(Sibling関係)、あるいは共通親関係(Co-Parent関係)、またはその両方を考慮したページの特徴抽出により関連ページ検索に適したページモデルを生成し、このページモデルに基づく関連ページ検索エンジンを提供することにより、関連ページの検索をより精度良く行われるようにすることを目的とする。

【００１０】

【課題を解決するための手段】

本発明の情報処理装置は、サイトを構成するページのデータを取得する取得手段と、取得手段により取得されたページのデータを用いて、ページ内に出現する単語を抽出する抽出手段と、抽出手段により抽出された単語が、ページ内で出現する回数をカウントするカウント手段と、取得されたページ間のリンク構造を解

析し、カウント手段によるカウントの値を用いて、リンク関係にあるページ間の第１の重みを生成する第１の生成手段と、第１の生成手段により生成された第１の重みを用いて、所定のページとリンク関係にある他のページ同士のページ間の第２の重みを生成する第２の生成手段と、第２の生成手段により生成された第２の重みを用いて、ＳＤＦ（Sibling Document Frequencyの略）データまたはＣＤＦ（Co-Parent Document Frequencyの略）データの少なくとも一方のデータを生成する第３の生成手段と、第３の生成手段により生成されたデータを用いて、ＩＳＤＦ（Inverse Sibling Document Frequencyの略）またはＩＣＤＦ（Inverse Co-Parent Document Frequencyの略）の少なくとも一方に基づくページモデル拡張処理により所定の値を算出する算出手段を含むことを特徴とする。

【００１１】

前記算出手段により算出された所定の値を用いて、取得されたページ内の所定のページ間の関連度を算出する第２の算出手段をさらに含むことを特徴とする。

【００１２】

前記第２の生成手段が、所定のページをリンク元とし、そのリンク元からリンクが張られているリンク先のページ同士のページ間の第２の重みを算出する場合、前記第３の生成手段は、ＳＤＦデータを生成し、前記算出手段は、ＩＳＤＦに基づくページモデル拡張処理により所定の値を算出し、前記第２の生成手段が、所定のページをリンク先とし、そのリンク先にリンクを張っているリンク元のページ同士のページ間の第２の重みを算出する場合、前記第３の生成手段は、ＣＤＦデータを生成し、前記算出手段は、ＩＣＤＦに基づくページモデル拡張処理により所定の値を算出し、前記第２の生成手段が、所定のページをリンク元とし、そのリンク元からリンクが張られているリンク先のページ同士のページ間の第２の重みと、所定のページをリンク先とし、そのリンク先にリンクを張っているリンク元のページ同士のページ間の第２の重みをそれぞれ算出する場合、前記第３の生成手段は、ＳＤＦデータとＣＤＦデータをそれぞれ生成し、前記算出手段は、ＩＳＤＦとＩＣＤＦに基づくページモデル拡張処理により所定の値を算出するようにすることができる。

【００１３】

前記算出手段は、所定のページ内における所定の単語の出現回数と、所定のページと前記第２の生成手段で生成されたリンク関係にあるページのうち、所定の単語を含むページに対応する前記第３の生成手段により生成されたデータを用いた演算により、所定の値を算出するようにすることができる。

【００１４】

前記第２の算出手段により算出された関連度を記憶する記憶手段と、所定のページに関連があるページの情報の提供が要求された場合、前記記憶手段に記憶されている関連度を参照して、所定のページと関連度が高いページの情報を提供する提供手段とをさらに含むようにすることができる。

【００１５】

前記提供手段は、情報を提供する際、所定のページと関連する広告に関する情報も提供するようにすることができる。

【００１６】

本発明の情報処理方法は、サイトを構成するページのデータを取得する取得ステップと、取得ステップの処理で取得されたページのデータを用いて、ページ内に出現する単語を抽出する抽出ステップと、抽出ステップの処理で抽出された単語が、ページ内で出現する回数をカウントするカウントステップと、取得されたページ間のリンク構造を解析し、カウントステップの処理によるカウントの値を用いて、リンク関係にあるページ間の第１の重みを生成する第１の生成ステップと、第１の生成ステップの処理で生成された第１の重みを用いて、所定のページとリンク関係にある他のページ同士のページ間の第２の重みを生成する第２の生成ステップと、第２の生成ステップの処理で生成された第２の重みを用いて、ＳＤＦデータまたはＣＤＦデータの少なくとも一方のデータを生成する第３の生成ステップと、第３の生成ステップの処理で生成されたデータを用いて、ＩＳＤＦまたはＩＣＤＦの少なくとも一方に基づくページモデル拡張処理により所定の値を算出する第１の算出ステップと、第１の算出ステップの処理で算出された所定の値を用いて、取得されたページ内の所定のページ間の関連度を算出する第２の算出ステップとを含むことを特徴とする。

【００１７】

本発明の記録媒体のプログラムは、サイトを構成するページのデータを取得する取得ステップと、取得ステップの処理で取得されたページのデータを用いて、ページ内に出現する単語を抽出する抽出ステップと、抽出ステップの処理で抽出された単語が、ページ内で出現する回数をカウントするカウントステップと、取得されたページ間のリンク構造を解析し、カウントステップの処理によるカウントの値を用いて、リンク関係にあるページ間の第１の重みを生成する第１の生成ステップと、第１の生成ステップの処理で生成された第１の重みを用いて、所定のページとリンク関係にある他のページ同士のページ間の第２の重みを生成する第２の生成ステップと、第２の生成ステップの処理で生成された第２の重みを用いて、ＳＤＦデータまたはＣＤＦデータの少なくとも一方のデータを生成する第３の生成ステップと、第３の生成ステップの処理で生成されたデータを用いて、ＩＳＤＦまたはＩＣＤＦの少なくとも一方に基づくページモデル拡張処理により所定の値を算出する第１の算出ステップと、第１の算出ステップの処理で算出された所定の値を用いて、取得されたページ内の所定のページ間の関連度を算出する第２の算出ステップとを含むことを特徴とする。

【００１８】

本発明のプログラムは、サイトを構成するページのデータを取得する取得ステップと、取得ステップの処理で取得されたページのデータを用いて、ページ内に出現する単語を抽出する抽出ステップと、抽出ステップの処理で抽出された単語が、ページ内で出現する回数をカウントするカウントステップと、取得されたページ間のリンク構造を解析し、カウントステップの処理によるカウントの値を用いて、リンク関係にあるページ間の第１の重みを生成する第１の生成ステップと、第１の生成ステップの処理で生成された第１の重みを用いて、所定のページとリンク関係にある他のページ同士のページ間の第２の重みを生成する第２の生成ステップと、第２の生成ステップの処理で生成された第２の重みを用いて、ＳＤＦデータまたはＣＤＦデータの少なくとも一方のデータを生成する第３の生成ステップと、第３の生成ステップの処理で生成されたデータを用いて、ＩＳＤＦまたはＩＣＤＦの少なくとも一方に基づくページモデル拡張処理により所定の値を算出する第１の算出ステップと、第１の算出ステップの処理で算出された所定の

値を用いて、取得されたページ内の所定のページ間の関連度を算出する第２の算出ステップとをコンピュータに実行させることを特徴とする。

【００１９】

本発明の情報処理装置および方法、並びにプログラムにおいては、ＩＳＤＦまたはＩＣＤＦの少なくとも一方に基づくページモデルにより、より精度の高い関連ページ検索が行われる。

【００２０】

【発明の実施の形態】

以下に、本発明の実施の形態について図面を参照して説明する。図１は、本発明の情報処理装置を含む情報処理システムの一実施の形態の構成を示す図である。ネットワーク１は、インターネットやＬＡＮ（Local Area Network）から構成されるネットワークである。ネットワーク１には、WWWサーバ２－１乃至２－３、端末３－１乃至３－３、および、検索サーバ４が接続され、相互にデータの授受を行えるように構成されている。

【００２１】

以下の説明において、WWWサーバ２－１乃至２－３を個々に区別する必要がない場合、単にWWWサーバ２と記述する。他の装置に関しても同様に記述する。なお、図１には、説明の都合上、WWWサーバ２や端末３は、それぞれ３台、検索サーバ４は１台しか図示していないが、それらの装置は、ネットワーク１に複数接続されている。

【００２２】

WWWサーバ２は、インターネット上のサービスの１つとして提供されているホームページを管理し、提供するサーバである。端末３は、ユーザ側の端末であり、WWWサーバ２から提供されるホームページを閲覧する機能を有する。検索サーバ４は、端末３のユーザが、WWWサーバ２で提供されるホームページに関連するページなどを検索したいときに接続されるサーバであり、ユーザの要求に対応する情報を検索し、その結果を提供する機能を有する。

【００２３】

図２は、WWWサーバ２の内部構成例を示す図である。WWWサーバ２は、パーソナ

ルコンピュータなどで構成することが可能であり、そのＣＰＵ（Central Processing Unit）１１は、ＲＯＭ（Read Only Memory）１２に記憶されているプログラムに従って各種の処理を実行する。ＲＡＭ（Random Access Memory）１３には、ＣＰＵ１１が各種の処理を実行する上において必要なデータやプログラムなどが適宜記憶される。入出力インタフェース１５は、キーボードやマウスから構成される入力部１６が接続され、入力部１６に入力された信号をＣＰＵ１１に出力する。また、入出力インタフェース１５には、ディスプレイやスピーカなどから構成される出力部１７も接続されている。

【００２４】

さらに、入出力インタフェース１５には、ハードディスクなどから構成される記憶部１８、および、ネットワーク１を介して他の装置（例えば、端末３）とデータの授受を行う通信部１９も接続されている。記憶部１８には、ホームページに関するデータが記憶されており、他の装置から、管理しているホームページの提供の要請があった場合に提供するようになされている。ドライブ２０は、磁気ディスク３１、光ディスク３２、光磁気ディスク３３、半導体メモリ３４などの記録媒体からデータを読み出したり、データを書き込んだりするときに用いられる。

【００２５】

図３は、端末３の内部構成例を示す図である。端末３は、パーソナルコンピュータなどで構成することが可能であり、そのＣＰＵ４１は、ＲＯＭ４２に記憶されているプログラムに従って各種の処理を実行する。ＲＡＭ４３には、ＣＰＵ４１が各種の処理を実行する上において必要なデータやプログラムなどが適宜記憶される。入出力インタフェース４５は、キーボードやマウスから構成される入力部４６が接続され、入力部４６に入力された信号をＣＰＵ４１に出力する。また、入出力インタフェース４５には、ディスプレイやスピーカなどから構成される出力部４７も接続されている。

【００２６】

さらに、入出力インタフェース４５には、ハードディスクなどから構成される記憶部４８、インターネットなどのネットワークを介して他の装置（例えば、検

索サーバ４）とデータの授受を行う通信部４９やドライブ５０も接続されている。記憶部４８には、WWWサーバ２から提供されるホームページを閲覧するために必要なブラウザなどのソフトウェアやデータが記憶されており、必要に応じ、読み出され、ＲＡＭ４３に展開される。

【００２７】

図４は、検索サーバ４の内部構成例を示す図である。検索サーバ４は、パーソナルコンピュータなどで構成することが可能であり、そのＣＰＵ７１は、ＲＯＭ７２に記憶されているプログラムに従って各種の処理を実行する。ＲＡＭ７３には、ＣＰＵ７１が各種の処理を実行する上において必要なデータやプログラムなどが適宜記憶される。入出力インタフェース７５は、キーボードやマウスから構成される入力部７６が接続され、入力部７６に入力された信号をＣＰＵ７１に出力する。また、入出力インタフェース７５には、ディスプレイやスピーカなどから構成される出力部７７も接続されている。

【００２８】

さらに、入出力インタフェース７５には、ハードディスクなどから構成される記憶部７８、インターネットなどのネットワークを介して他の装置（例えば、端末３）とデータの授受を行う通信部７９やドライブ８０も接続されている。記憶部７８には、WWWサーバ２により提供されるホームページを検索するためのデータが記憶されている。

【００２９】

図５は、検索サーバ４の機能ブロック図である。検索サーバ４は、データを記憶する記憶機能と、その記憶されるデータを作成したり、記憶されているデータを用いた処理を実行する処理機能とから構成されている。検索サーバ４は、記憶機能として、データを収集するホームページ（サイト）のリストを記憶する収集サイトリスト記憶部１０１、収集サイトリスト記憶部１０１に記憶されているリストに基づき収集されたサイトのページのデータを記憶する保存ページ記憶部１０２、および、保存ページ記憶部１０２に記憶されたページデータが処理された結果を記憶するページデータ記憶部１０３を備えている。

【００３０】

検索サーバ４は、処理機能として、保存ページ記憶部１０２に記憶されているページデータを処理するサイトページ処理部１１１と、サイトページ処理部１１１により処理された結果としてのデータを用いて所定の処理を実行し、関連ページに関するデータの生成などを行う関連ページデータ処理部１１２を備えている。

【００３１】

サイトページ処理部１１１により処理されたデータは、ページデータ記憶部１０３のサイトページデータ記憶部１０４に記憶され、関連ページデータ処理部１１２により処理されたデータは、ページデータ記憶部１０３の関連ページデータ記憶部１０５に記憶される。

【００３２】

サイトページ処理部１１１およびサイトページデータ記憶部１０４の詳細について、図６を参照して説明する。サイトページ処理部１１１は、ページ取得保存部１４１を備える。ページ取得保存部１４１は、収集サイトリスト記憶部１０１に記憶されているリストに記載されているサイトと接続する処理を実行し、各々のサイトに記憶されているホームページの全てのページのデータをダウンロードし、そのダウンロードしたデータを保存ページ記憶部１０２に記憶（保存）させる。

【００３３】

保存ページ記憶部１０２に記憶されたページは、ページＩＤ割り当て部１４２により、各ページが一意に区別がつくようなＩＤが割り当てられ、その割り当てられたＩＤに関するデータが、サイトページデータ記憶部１０４のページＩＤ記憶部１６１に記憶される。

【００３４】

保存ページ記憶部１０２に記憶されたページは、単語抽出部１４３にも読み出される。単語抽出部１４３は、読み出したページ内から、そのページに含まれる単語を抽出する。単語抽出部１４３により抽出された単語のデータは、単語ＩＤ割り当て部１４４に供給される。単語ＩＤ割り当て部１４４は、供給された単語に対して、その単語が他の単語と区別がつくようなＩＤを割り振る。その割り振

られたＩＤと、そのＩＤに対応する単語のデータは、サイトページデータ記憶部１０４の単語ＩＤ記憶部１６２に記憶される。

【００３５】

単語割り当て部１４４からのデータは、基本ページモデル生成部１４５にも提供される。基本ページモデル生成部１４５は、抽出された単語が、そのページ内で、どのぐらいの頻度で用いられているかなどのデータを作成する。基本ページモデル生成部１４５により作成されたデータは、サイトページデータ記憶部１０４の基本ページモデル記憶部１６３に記憶される。

【００３６】

保存ページ記憶部１０２に記憶されているページはサイトページ処理部１１１のリンク判定部１４６にも読み出される。リンク判定部１４６は、各ページの親子関係を判定する。各ページの親子関係とは、所定のページにおいて、そのページを親ページと称したとき、その親ページがリンクを張っている先のページを子ページと称したときの関係である。リンク判定部１４６により判定されたページ間の親子関係に関する情報は、サイトページデータ記憶部１０４のリンク情報記憶部１６４に出力され、記憶される。

【００３７】

次に、関連ページ処理部１１２と関連ページデータ記憶部１０５の詳細な構成について、図７を参照して説明する。関連ページ処理部１１２は、必要に応じ、サイトページ記憶部１０４に記憶されているデータを用いて処理を実行する。まず、関連ページ処理部１１２のリンク関係情報生成部１８１は、サイトページデータ記憶部１０４に記憶されているデータを用いて同じ親ページを持つ子ページの情報を抽出する。

【００３８】

図８を参照して説明するに、１つ所定の親ページからリンクが張られている子ページが複数存在している場合、その子ページの情報が抽出される。そして抽出された子ページ同士の情報、すなわち、兄弟(Sibling)と位置づけられるページ同士の情報が生成される。兄弟と位置づけられるページ同士の情報が、リンク関係情報生成部１８１において生成され、関連ページデータ記憶部１０５のリンク

関係情報記憶部１９１に記憶される。

【００３９】

関連ページ処理部１１２のＳＤＦデータ生成部１８２は、ＳＤＦデータを生成する。ＳＤＦとは、Sibling Document Frequencyの略である。ＳＤＦデータ生成部１８２により生成されるＳＤＦデータとは、詳細は後述するが、各々のページに含まれる（各ページで出現する）単語に対して、その単語が現れる兄弟ページのリンクの重みを総和したデータである。

【００４０】

ＳＤＦデータ生成部１８２により生成されたＳＤＦデータは、関連ページデータ記憶部１０５のＳＤＦデータ記憶部１９２に記憶される。関連ページ処理部１１２のページモデル拡張部１８３は、ＳＤＦデータ記憶部１９２に記憶されているデータに対して重み付けを行い、その重み付けを行ったデータを、関連ページデータ記憶部１０５のページモデル拡張データ記憶部１９３に提供し、記憶させる。

【００４１】

関連ページ処理部１１２の関連度算出部１８４は、ページ毎の関連度を算出し、その結果を、関連ページデータ記憶部１０５の関連度データ記憶部１９４に記憶させる。関連度算出部１８４が行う関連度の算出は、例えば、ＶＳＭ（Vector Space Modelの略、あるいは、ベクトル空間法と称される）のcosine類似度に基づいて行われる。

【００４２】

関連ページ一覧生成部１８５は、ユーザからの指示があった場合に、ページデータ記憶部１０３に記憶されているデータに基づいて、関連するページについての一覧表を作成し、そのデータを提供するといった処理を実行する。

【００４３】

このようなデータを生成し、記憶する検索サーバ４と端末３との間で行われる処理について、図９のフローチャートを参照して説明する。ステップＳ１１において、端末３は、ネットワーク１を介して検索サーバ４に接続される。ここでの接続（アクセス）は、始めて端末３が検索サーバ４と接続されたとき、又は、端

末３側で後述する設定が行われていないときであるとする。換言すれば、ユーザが検索を行うために、後述する関連ページ検索ボタン２３１（図１０Ｂ）を操作したときの接続とは異なる接続であるとする。

【００４４】

検索サーバ４は、端末３からのアクセスを受け付けると、ステップＳ２１において、導入画面の送付を行う。導入画面とは、端末３のユーザが、検索サーバ４による検索を行う際に操作するボタンなどを、端末３のブラウザ上に設定するための画面であり、例えば、図１０Ａに示したような画面である。

【００４５】

端末３の記憶部４８（図３）には、ネットワーク１を介してデータの授受を行う際に用いられるブラウザに関するプログラムが記憶されており、必要に応じ、起動され、ＣＰＵ４１が処理を実行する際に用いられる。ブラウザが起動され、検索サーバ４からの導入画面のデータが受信され、起動されているブラウザにより処理されると、図１０Ａに示したような画面が、出力部４７としてのディスプレイ２１１上に表示される（ステップＳ１２）。

【００４６】

ディスプレイ２１１には、ブラウザが起動されることにより表示される部分の下側に、画像表示部２２１が設けられており、その画像表示部２２１に、検索サーバ４からの導入画面が表示される。導入画面としては、例えば、“このボタンをドラッグアンドドロップすると、関連ページ検索エンジンがブラウザ上に設定されます”といったメッセージと共に、ボタンが表示されている画面である。ユーザは、このメッセージに従って、ボタンを、例えば、ブラウザの上部の所定の欄（通常、リンクツールバーという欄）にドラッグアンドドロップを行う。

【００４７】

このようなドラッグアンドドロップが、ステップＳ１３において行われると、そのドラッグアンドドロップの処理に対応する設定が、ステップＳ１４において、行われる。すなわち、例えば、図１０Ｂに示したように、ドラッグアンドドロップされたボタンに対応する関連ページ検索ボタン２３１が、ブラウザの所定の部分に表示され、その関連ページ検索ボタン２３１に関連付けられて、検索サー

バ４のアドレスが記憶されるなどの設定である。

【００４８】

このような設定が行われることにより、図１０Ｂに示したように、ブラウザ上の所定の部分に関連ページ検索ボタン２３１が表示されるようになると、ユーザは、検索サーバ４による検索を利用することが可能な状態とされる。

【００４９】

このような導入画面を用い、関連ページ検索ボタン２３１がブラウザ上に設定されるようにしても良いし、所定のページ内に、バナーとして関連ページ検索ボタン２３１が設けられているようにしても良い。また、ユーザが検索サーバ４にアクセスし、所定のページのＵＲＬ（Uniform Resource Locationの略）を入力することも可能である。いずれにしても、ユーザが検索を所望したときに、ボタンのクリックなど簡便な操作で、検索サーバ４にアクセスでき、その検索サーバ４による検索の結果を授受できるように設定されていればよい。

【００５０】

ここでは、図１０Ｂに示したように、ブラウザ上に、関連ページ検索ボタン２３１が設定されているとして説明する。ユーザが、所定のページ、例えば、ＷＷＷサーバ２－１（図１）により提供されているホームページの所定のページを閲覧している状態で、関連ページ検索ボタン２３１を操作すると、検索サーバ４に、関連ページ検索ボタン２３１が操作されたという情報、すなわち、検索が指示されたという情報が送信される。その結果、検索サーバ４においては、図１１に示すようなフローチャートの処理が開始される。

【００５１】

ステップＳ４１において、所定のホームページ（サイト）のページのデータ（以下、単にページと記述した場合も、ページのデータという意味を示すとする）が取得され、保存される。取得されるホームページのページは、収集サイトリスト記憶部１０１に記憶されているリストに基づくものである。また、ユーザからの要求により送信された所定のＵＲＬが収集サイトリスト記憶部１０１に記録されていない場合には、そのＵＲＬを追加し記録する。収集サイトリスト記憶部１０１に記憶されているリストの一例を図１２に示す。図１２に示したように、収

集サイトリスト記憶部１０１に記憶されているリストには、“収集開始ＵＲＬ”、“含むディレクトリ”、“排他ディレクトリ”、“含むドメイン”、および、“排他ドメイン”といった情報が含まれる。

【００５２】

このようなリストに基づいて、ページが取得される。取得されたページは、保存ページ記憶部１０２に保存され記憶される。また、保存ページ記憶部１０２には、図１３に示すようなリスト形式で、取得されたページのサイト単位での情報が管理されている。図１３に示したように、リストには、“サイトＩＤ”、“サイト名”、および、“総ページ数”といった情報が含まれる。

【００５３】

サイトＩＤは、そのサイトに割り当てられたＩＤであり、ページ取得保存部１４１がページ（サイト）の情報を取得した時点で、割り当てるようにしても良いし、収集サイトリスト記憶部１０１で記憶されている図１２に示したようなリストで、ＩＤも関連付けて記憶するようにし、そのＩＤが、記憶されるようにしても良い。

【００５４】

このようにして、保存ページ記憶部１０２に取得されたページが保存され、所定のサイトの情報が記憶されると、ステップＳ４２において、ページＩＤ割り当て部１４２により、取得されたページ毎にＩＤが割り当てられる。ページＩＤ割り当て部１４２は、保存ページ記憶部１０２に記憶されているページを読み出し、そのページにＩＤを割り当てる。

【００５５】

この際、読み出されるページと、割り当てられたＩＤから、図１４に示したようなリストが作成され、ページＩＤ記憶部１６１に記憶される。図１４に示したページＩＤ記憶部１６１に記憶されるリストには、“ページＩＤ”、“サイトＩＤ”、“ページＵＲＬ”、“タイトル”、“サマリー”、“ページ保存場所”、および、“最終更新日”といった情報が含まれる。

【００５６】

これらの情報のうち、“ページＩＤ”は、ページ割り当て部１４２により割り

当てられ、その他の情報は、保存ページ記憶部１０２に記憶され、読み出されたページのデータから抽出される。

【００５７】

ステップＳ４３において、ページ内に含まれる単語が、単語抽出部１４３により抽出される。この単語の抽出は、単語抽出部１４３が、保存ページ記憶部１０２から保存されているページのうちの１ページを読み出し、そのページに含まれている単語が抽出されることにより行われる。抽出される単語としては、名詞に分類される単語である。形容詞や動詞などに分類される単語や、英語なども抽出されるようにしても良い。単語抽出部１４３により抽出される単語は、後の処理において必要となる（検索サーバ４が最終的に検索結果としてユーザに提供する結果が良好になるために必要となる）品詞の単語が抽出されれば良い。

【００５８】

抽出された単語は、単語ＩＤ割り当て部１４４に供給される。単語ＩＤ割り当て部１４４に供給されるのは、抽出された単語だけでなく、その単語の出現回数、ページＩＤ、タグ付きの各単語、そのタグ付きの単語の出現回数なども供給される。これらの情報は、必要に応じ、単語抽出部１４３が、ページＩＤ記憶部１６１や保存ページ記憶部１０２から読み出し、供給する。

【００５９】

単語ＩＤ割り当て部１４４は、供給された単語に対してＩＤを割り当てる。ＩＤが割り当てられた単語は、ＩＤと関連付けられて、単語ＩＤ記憶部１６２に記憶される。単語ＩＤ記憶部１６２には、例えば、図１５に示したようなリストが記憶されている。

【００６０】

図１５に示したように、単語ＩＤ記憶部１６２には、“単語ＩＤ”と“単語”が関連付けられて記憶されている。なお、同一の単語が抽出された場合には、同一のＩＤが割り当てられる。そのために、単語抽出部１４３は、抽出された単語が、既に単語ＩＤ記憶部１６２に記憶されている単語であるか否かを判断し、既に記憶されている単語である場合には、新たにＩＤを割り振ることが無いように制御されている。

【００６１】

また、単語ＩＤ割り当て部１４４は、図１６に示したようなリストを作成し、単語ＩＤ記憶部１６２に記憶させる。図１６に示したリストは、“単語ＩＤ”、“サイトＩＤ”、“そのサイト内で当該単語を含むページ数”、および、“そのサイト内で当該単語を含むページ”といった情報を含むものである。図１６に示したリストは、所定の１つのサイトに注目したときに、そのサイトに含まれる所定の単語との関連を示すものである。

【００６２】

単語ＩＤと、その単語ＩＤが割り当てられた単語は、単語ＩＤ記憶部１６２に供給されると共に、その一部のデータは、基本ページモデル生成部１４５にも供給される。基本ページモデル生成部１４５は、ステップＳ４５において、基本ページモデルを生成する。基本ページモデルとは、図１７に示したようなデータであり、基本ページモデル記憶部１６３に記憶されるリスト形式のデータである。このようなデータを作成するために、基本ページモデル生成部１４５は、単語ＩＤ割り当て部１４４から、ページＩＤと、それぞれの単語ＩＤとその出現回数に関する情報が供給される。

【００６３】

図１７に示すように、基本ページモデル記憶部１６３に記憶されるリストは、“ページＩＤ”、“出現単語”、“Title”、“Keywords”、および、“description”といった情報が含まれる。このリストは、１つのページに対して、１つの単語が、何回出現しているか（用いられているか）を示す情報であり、また、タイトル（Title）などの種類毎に分類された情報も含む情報である。このような種類毎に分類された情報は、最終的に関連するページを決定する際に、単語の用いられている部分毎（種類毎）に重要度が異なることが考えられ、そのような重要度の違いにより重み付けを行うようにした場合のことを考慮したときに用いられる。

【００６４】

ステップＳ４６において、リンク判定部１４６は、図８を参照して説明したように、親ページと、そのページがリンクしている子ページを判断し、その判断結

果をリンク情報記憶部１６４に記憶させる。リンク情報記憶部１６４に記憶されている情報は、例えば、図１８に示したような情報である。

【００６５】

図１８に示したように、リンク情報記憶部１６４に記憶されているリスト形式の情報は、“ページＩＤ”、“リンク先のページＩＤ”、“リンクの重み”、および、“アンカー窓内単語”といった情報が含まれている。“ページＩＤ”と、“リンク先のページＩＤ”、すなわち、親としてのページと子としてのページが関連付けられていることがわかる。このような情報を作成するために、リンク判定部１４６は、必要に応じ、保存ページ記憶部１０２、ページＩＤ記憶部１６１、および、基本ページモデル記憶部１６３からデータを読み出す。

【００６６】

“リンクの重み”は、以下のようにして算出される。なお、重み付けを算出する際、アンカー窓内に含まれる単語が、リンク先のページ（この場合、子ページ）に含まれるほど、ページ間の関連性が高いと考え、重みを増加するようにする。また、リンク元のページ（すなわち、親ページ）が多くのリンクを持つほど、１つのリンクに対する重要性は低いと考えられるため、そのようなページからリンクが張られている子ページとのリンクの重みは小さくなるようにする。

【００６７】

親ページpから子ページqへのリンクの重みWc（p,q）は、次式（１）に基づいて算出される。

$$Wc(p,q) = 1 + Npq(Tanc) \times 1/k \quad \cdots (1)$$

式（１）において、p,q∈Ｐ（Ｐはページ集合）である。また、Npq（Tanc）は、親ページp内のアンカー窓内の単語の集合を集合（Tanc）とし、その集合（Tanc）の子ページq内の出現数を表している。なお、Tanc∈Tallであり、Tallは、全単語の集合とする。

【００６８】

ｋは、親ページｐが有しているリンクの数であり、ページpからページqへのリンクを含むため、ｋは、常に１以上の数に設定される。なお、式（１）において、右辺の第１項で１だけ加算しているのは、算出される重みWc（ｐ，ｑ）が１

未満にならないようにするためである。

【００６９】

このようにして重みWc（ｐ，ｑ）が算出されるようにしても良いし、アンカー窓内の出現単語にアンカーを中心とした距離に応じた重み付けを行ってWc（ｐ，ｑ）が算出されるようにしても良い。アンカーを中心とした距離に応じた重み付けを行ってWc（ｐ，ｑ）を算出するようにした場合、式（１）におけるNpq（Tanc）は、次式（２）に基づいて算出される。

$$N(p,q)(Tanc)=H(Dis(t1))\times Tc(t1)+H(Dis(t2))\times Tc(t2)+\cdots+H(Dis(tk))\times Tc(tk) \quad \cdots (2)$$

【００７０】

式（２）において、tk∈Tancであり、Dis（tk）は、アンカータグから単語tkが出現するまでの距離を示し、　０≦Dis（tk）≦Dmax の値を取る。Dmaxは、アンカー窓の片方の最大幅である。また、Ｈ（Dis（tk））は、Dis（tk）に対する重みを表し、　０＜ Ｈ（Dis（tk））　≦１ の範囲内の値であり、Ｈ（0）＝１である。Tc（tk）は、単語tkの子ページq内の出現数を表す。

【００７１】

このように、アンカー窓からの距離を考慮した重み付けを行うようにしても良い。また、アンカー窓内の単語のアンカー窓内出現数や、リンク先ページ（子ページ）での出現数にタグの種類に応じた重み付け（重要度）を考慮した重み付けを行うことも可能である。また、これらの重み付けを行わず、単にWc（p,q）＝１としてもよい。

【００７２】

このようにして、図１８に示したリンク情報記憶部１６４に記憶されるリスト内の“リンクの重み”は算出される。図１１のフローチャートの説明に戻り、ステップＳ４７において、リンク関係情報の生成が、リンク関係情報生成部１８１（図７）により行われる。リンク関係情報生成部１８１により作成された情報は、リンク関係情報記憶部１９１（図７）に、図１９に示したようなリスト形式で記憶される。リンク関係情報生成部１８１は、図１９に示したような情報を作成するための情報を、リンク情報記憶部１６４から取得する。

【００７３】

図１９に示したように、リンク関係情報記憶部１９１には、“ページＩＤ”、“Siblingページ ＩＤ”、および、“リンクの重み”が、それぞれ関連付けられて記憶されている。ここで、Siblingページとは、共通の親ページを有する子ページのことであり、図８を用いて説明したように、兄弟関係にあるページのことを示す。

【００７４】

リンク関係情報生成部１８１は、各々のページＩＤに対して、Siblingの関係にあるページＩＤを抽出するといった処理を行うとともに、Siblingページ間のリンクの重みも算出する。そのSiblingページ間のリンクの重みの算出は、以下のようにして行われる。すなわち、Siblingページ間のリンクの重みWs（r,s）は、次式（３）に基づいて算出される。

【００７５】

$$Ws(r,s) = Wc(t,r) \times Wc(t,s) \quad \cdots (3)$$

式（３）において、r,s,tは、Ｐをページ集合とした場合、r,s,t∈Ｐを満たす値であり、Ws（r,s）は、１≦Ｗｓ（r,s）を満たす値である。

【００７６】

式（３）において、Ws（r,s）は、所定のページrと、そのページrとSiblingの関係にあるSiblingページs間のリンクの重みであり、Wc（t,r）は、所定のページtと、そのページtと親子関係にある子ページr間のリンクの重みであり、Wc（t,s）は、所定のページtと、そのページtと親子関係にある子ページs間のリンクの重みである。

【００７７】

図２０を参照して、式（３）について説明するに、この重みの算出は、所定のページrと、そのページrとSibling（兄弟）の関係にあるページsとのリンクの重みWs（r,s）は、そのSibling関係内に存在するページ間のリンクの重み、この場合、ページrとページsとに共通に親子の関係にある親ページtとのリンクの重みである、重みWc（t,r）と重みWc（t,s）とを乗算することにより求められる。

【００７８】

このようにして、Siblingページ間のリンクの重みが算出され、その算出結果が、図１９に示したようなリスト形式のデータに書き込まれる。

【００７９】

図１１のフローチャートの説明に戻り、ステップＳ４８において、ＳＤＦデータの生成が、ＳＤＦデータ生成部１８２（図７）により行われる。ＳＤＦデータ生成部１８２は、必要に応じ、リンク関係情報記憶部１９１と基本ページモデル記憶部１６３からデータを読み出し、その読み出したデータを用いて、図２１に示すようなリスト形式のデータを作成し、ＳＤＦデータ記憶部１９２に記憶させる。

【００８０】

図２１に示したＳＤＦデータ記憶部１９２に記憶されるデータは、“ページＩＤ”と、“ページＩＤに含まれる単語ＩＤと、その単語ＩＤを含むSiblingページのリンクの重みの総和”といった情報を含む。このデータは、各々のページにおいて、そのページ内で出現する単語に対して、その単語が現れるSiblingページのリンクの重みを総和したデータであり、リンク判定部１４６が、前述のとおり$Wc(p,q)=1$とリンクの重みを生成した場合には、単にその単語が現れるSiblingページの総数となる。

【００８１】

ステップＳ４９において、ページモデル拡張部１８３（図７）は、ページモデル拡張処理を実行する。ページモデル拡張処理とは、図２２に示すようなリスト形式のデータを作成し、ページモデル拡張データ記憶部１９３に記憶させる処理である。ページモデル拡張部１８３は、図２２に示すようなデータを作成するために、基本ページモデル記憶部１６３、リンク情報記憶部１６４、リンク関係情報記憶部１９１、および、SDFデータ記憶部１９２に、それぞれ記憶されているデータを必要に応じて読み出す。

【００８２】

図２２に示したページモデル拡張データ記憶部１９３に記憶されているデータは、“ページＩＤ”と“ベクトル”といった情報を含む。“ベクトル”内の重み

は、ＩＳＤＦ（Inverse Sibling Document Frequency）に基づき、以下のようにして求められる。

【００８３】

Ｐｉ＝(|Ti1×Wi1|, |Ti2×Wi2|, ……, |Tij×Wij|, ……　・・・（４）

式（４）において、iはページであり、i∈Ｐ、jは単語であり、j∈Tallである。Ｐiは、ページiのTall次元のベクトルを示す。Tijは、ページiにおいて単語jが出現しているか否かを示す値であり、出現している場合１が、出現していない場合０が、それぞれ設定される。

【００８４】

Wijは、ページiにおける単語jの重みであり次式（５）に基づいて算出される。また、Wij は、 0 ≦ Wij を満たす値となり、Σ(Ti×Wij)＾２＝１（TiとWijを乗算した値を２乗した値の総和が１）になるよう正規化される。

【００８５】

Wij=（１＋log(TFij)）×（１＋log(１/(１＋SDFij)))　・・・（５）

式（５）において、TFij は、単語jのページiにおける出現回数を示し、0 ≦ TFijの値を取る。SDFij は、ページiのSiblingページのうち、単語jを含むページのリンクの重みの総和を示す。

【００８６】

このような式（４）と式（５）を用いてベクトル内の重みを算出するようにしても良いが、さらに、SDFijの効果を高めるため、式（５）を式（６）に置き換えても良い。

Wij = （１＋log(TFij)）×（１＋log(１+ASDFi /(１＋SDFij)))

・・・（６）

【００８７】

式（６）において、ASDFiは、ページiと全Siblingページ間のリンクの重みの総和を示す。

【００８８】

さらに、TTFij, ATFijを加え、式（５）を基に次式（７）あるいは、式（６）を基に次式（８）に基づいて、重みを算出するようにしても良い。

$$Wij = (1 + \log(TFij+TTFij+ATFij)) \times (1 + \log(1/(1 + SDFij))) \quad \cdots (7)$$

$$Wij = (1 + \log(TFij+TTFij+ATFij)) \times (1 + \log(1+ ASDFi/(1 + SDFij))) \quad \cdots (8)$$

式（７），（８）において、TTFijは、タグ付単語jがページiにおいて出現するか否かを示し、出現しない場合０が、出現する場合１が、それぞれ設定される。あるいは出現回数(0以上)が設定されるようにしても良い。タグの種類に応じてそれぞれ重みを付けるようにしても良い。

【００８９】

また、ATFijは、ページiのリンク元ページ（この場合、親ページ）におけるアンカー窓内において単語jが出現するか否かを示し、出現しない場合0が、出現する場合１が、それぞれ設定される。あるいは出現回数(0以上)が設定されるようにしても良い。タグ付単語と同様に、重みを付けてもよい。さらに、アンカーからの距離に応じて重みをつけるようにしても良い。

【００９０】

このような式に基づいて、図２２に示したデータ内の“ベクトル”内の各々の単語に対する“重み”のデータが算出される。図１１のフローチャートの説明に戻り、ステップＳ５０において、関連度算出部１８４において、ページ間の関連度が算出される。関連度算出部１８４は、必要に応じ、ページモデル拡張データ記憶部１９３に記憶されているデータを読み出し、図２３に示すようなリスト形式のデータを作成し、関連度データ記憶部１９４に記憶させる。

【００９１】

図２３に示した関連度データ記憶部１９４に記憶されるデータは、“ページＩＤ”、“対象ページＩＤ”、“関連度”、および、“高関連度単語”といった情報を含む。このうち、関連度は、以下のようにして算出される。関連度は、関連ページ検索に適した形に特徴抽出されたページ間の特徴が共通している部分が多いほど、関連度が高くなるという考えに基づき算出される。例えば、共通特徴数／総特徴数（積／和）、ＶＳＭのcosine類似度などを用いて算出することができる。

【００９２】

具体的には、次式（９）に基づいて算出される。式（９）は、ＶＳＭのcosine類似度によるものである。

$$R(i,j) = Pi \cdot Pj / | |Pi| | |Pj| | \quad \cdots (9)$$

式（９）において、Pi 、Pjはそれぞれページi、ページｊのベクトル表現であり、式（４）により算出（表現）される値である。また、i、j∈Ｐである。Ｒ(i,j)は、ページｉに対するページjの関連度であり、図２３では、ページiが“ページＩＤ”，ページjが“対象ページＩＤ”となる。

【００９３】

このようにして算出された関連度は、図２３に示すようなリスト形式のデータ内のデータとして、関連度データ記憶部１９４に記憶される。次に、ステップＳ５１以降の処理が行われるわけだが、ステップＳ５１以降の処理は、このようにして各記憶部に記憶されたデータ、特に、関連度データ記憶部１９４に記憶されているデータが用いられて行われる。

【００９４】

そこで、ここまでの処理、すなわち、ステップＳ４１乃至Ｓ５０までの処理は、ユーザの要求があった際に、リアルタイムに実行されるようにしても良いし、ユーザの要求に関わらず、事前に実行されるようにしても良い。

【００９５】

ユーザの要求に関わらず、ステップＳ４１乃至Ｓ５０の処理が行われる場合、所定のサイトから定期的にデータを取得するようにし、各記憶部に記憶されているデータが更新されるようにすれば良い。このように、予めデータを作成しておけば、ユーザからの要求があった際、ユーザからの要求があってからリアルタイムに処理を実行するよりも、その要求に即座に対応することが可能となる。

【００９６】

また、上記のように予めデータを作成した場合、ユーザから要求がある際に送信されるＵＲＬが予め作成したデータに存在しないときには、ステップＳ４１乃至Ｓ５０をそのＵＲＬの示すページ、あるいはそのページのサイトについて行うことが可能である。

【００９７】

ステップＳ５１において、関連ページ一覧生成部１８５は、ユーザが関連ページの提供を指示してきたページに対応する関連ページの一覧を作成する。その作成は、以下のようにして行われる。

【００９８】

まず、関連ページ一覧生成部１８５は、ページＩＤ記憶部１６１から、ユーザが関連ページ検索ボタン２３１を操作した際に閲覧されていたページ（関連ページの検索が指示されたページ）のＵＲＬに対応するページＩＤを読み出す。その読み出されたページＩＤをＫｅｙ１とするデータが、関連度データ記憶部１９４（図２３）から読み出される。その際、関連度の値が高い順にソートされ、その関連度に該当する対象ページＩＤ（Ｋｅｙ２となるページＩＤ）が読み出される。

【００９９】

そして、関連ページ一覧生成部１８５は、該当したページＩＤをページＩＤ記憶部１６１に照合し、ＵＲＬなど、そのページに関する情報を取り出し、一覧データを生成する。

【０１００】

一覧データを生成する際、ここまでの処理により得られたデータで終了しても良いが、さらに、以下のような機能を付け加えても良い。ユーザには、関連度が高い順にページに関する情報が表示されるように、一覧表が作成されるわけだが、例えば、同一の関連度を有するページが複数存在する場合が考えられ、そのようなとき、どのページを上位に表示するかが問題となる。また、関連度とは関係しないページの重要度を加味して、最終的にユーザへ関連ページを表示することも考えられる。

【０１０１】

そこで、関連度算出部１８４が算出した関連度に対して、ページのランク付けを行い、そのデータを最終的な関連度の値に付加するようにする。例えば、ページのランク付けとしては、検索サーバ４自体が、ランク付けの機能を有するようにしても良いし、他のサーバで提供しているランク付けの情報を引用するように

しても良い。

【０１０２】

ランク付けのデータを加味した関連度の算出は、具体的には、パラメータによる調整が考えられる。

$$R'(i,j) = p\,R(i,j) + (1-p)\,G(j) \quad \cdots \quad (10)$$

式（１０）において、R’（i,j）は、ページiに対するページjのランク付関連度であり、R(i,j)は、ページiに対するページjの関連度であり、式（９）により算出される値である。また、G(j)は、ページjのランクであり、ｐは、０≦ｐ≦１の値を有するパラメータである。この式（１０）で算出されたランク付関連度をすでに述べた図２３に示すようなリスト形式のデータ内のデータとして、関連度データ記憶部１９４に記憶してもよい。

【０１０３】

また、上述した実施の形態でステップＳ４９において、ページモデル拡張部１８３が行う処理の前または後の処理として、リンク先のページを考慮したページモデルを作成するようにしても良い。具体的には、所定のページの基本ページモデルに、リンク先の基本ページモデルの総和を付加する。このようにした場合、上述したリンク判定部１４６で算出されるリンク間の重みを付加するようにしてもよい。最下層（葉）のページまで計算する、あるいは、Ｎ回のリンク先まで考慮という形にする。

【０１０４】

ISDFによるページモデル拡張部１８３が行う処理の前に、この機能を実現した場合、所定のページのページモデルに存在する単語種が増えるため、ISDFの結果が影響を受けることになるため、このことを考慮して、前または後の、どちらに処理を実行するかを決定した方が良い。

【０１０５】

さらに、上述した実施の形態において、各処理を行う上で、単語の関連性ということを考慮して処理を行うようにしても良い。例えば、“旅行”と“海外”といった単語を関連付けた辞書（関連辞書）を設け、その関連辞書を参照して処理が行われるようにする。このような関連辞書を設けない場合は、ページ内に出現

した単語のみで関連度が決定されるが、関連辞書を設けるようにした場合は、例えば、基本ページモデル生成部１４５やＳＤＦデータ生成部１８２、あるいは関連度算出部１８４などが処理を実行する前の処理として、関連辞書が参照され、その結果が用いられて関連度が算出されるようにしても良い。関連辞書としては、共起情報やKeyGraph手法により作成されるか、ＯＤＰ（Open Directory Projectの略）のカテゴリー情報などが利用されるようにしても良い。

【０１０６】

図１１のフローチャートの説明に戻り、このようにして生成された一覧データは、ステップＳ５２において、ネットワーク１を介して端末３に送信される。端末３側において、一覧データが処理されることにより、ユーザに関連ページの一覧表が提供される。この関連ページの一覧表は、端末３のディスプレイ２１１上では、既に開かれているウインドウ（関連ページ検索ボタン２３１が操作されたウインドウ）とは異なるウインドウとして表示されるようにしても良いし、既に開かれているウインドウに表示されるようにしても良い。

【０１０７】

ここで、このような検索サーバ４による検索の結果として、ユーザに提供される関連ページについて説明する。例えば、従来の手法により所定のページの関連ページを検索した場合、その検索される関連ページは、類似しているページが上位に表示されるようになっていた。例えば、所定のミュージシャンのサイト内のプロフィールのページを閲覧しているときに、そのページに関連するページを検索した場合、そのミュージシャンの他のサイト内のプロフィールのページが検索結果としてユーザに提供されるといったことが行われていた。

【０１０８】

しかしながら、この例の場合、同一のミュージシャンの同一のプロフィールを、別のサイトで閲覧してもユーザにとって、新たに得られる情報は何もないといえる。換言すれば、ユーザは、同一のミュージシャンのプロフィールを何度も閲覧したいわけではなく、プロフィールに関連する情報、例えば過去に参加したイベントに関する情報や、プロフィールに記載されたストーリーに関する情報、ミュージシャンが好む事柄に関する情報などを所望しているために、関連ページの

検索を実行したと考えられる。すなわち、ユーザは、検索を実行する際、重複した情報である類似するページを参照したいわけではなく、何らかの関わりのあるページを参照したいと考えられる。このような、類似しているわけではないが、関連しているページを提供することが、上述した検索サーバ４による検索においては実現することが可能である。

【０１０９】

上述した検索サーバ４の処理を図２４を参照して説明する。図２４に示すように、親ページには、リンクが張られている子ページとして子ページ１乃至３が存在するとする。そして、子ページ１に含まれる単語（ステップＳ４３の処理で抽出される単語）が、“ａ，ｂ，ｃ，・・・”であり、子ページ２に含まれる単語が、“ａ，ｃ，ｄ，・・・”であり、子ページ３に含まれる単語が、“ａ，ｘ，・・・”であるとする。

【０１１０】

このような状況では、子ページ１乃至３には、共通に、単語ａが含まれている。例えば、所定の会社が運営するサイトの所定の製品Ａのホームページ内で、使い方の提案などが掲載されているページがあるとする。そのページ内には、製品Ａの名称を示す単語ａが、高い確率で含まれている可能性がある。そのような場合には、単語ａは、各ページの特徴を示す単語として（他のページとの差異を表す単語として）は、ふさわしくないと考えられる。

【０１１１】

よって、単語ａなど、複数のページに共通に含まれる単語などは、それらのページの特徴を表す単語として取り扱われないようにする。換言すれば、ページ間の関連度を判断するためのページの特徴抽出としては、単語ａなど、複数のページに共通に含まれる単語などは、他の単語と比較して重要度が低く設定される（他の単語の方が、重みが重く設定される）ようにする。

【０１１２】

その重みの設定は、上述したように、本実施の形態においては、ＩＳＤＦ（Inverse Sibling Document Frequency）に基づいて行っている。このＩＳＤＦに基づく重み付けは、上述したように、ステップＳ４９の処理として、ページモデル

拡張部１８３（図７）が行っている。

【０１１３】

ここで、従来の重み付けの手法として、ＴＦ－ＩＤＦ（Term Frequency-Inverse Document Frequency）がある。重み付けにＴＦを用いるのは、文書中（所定のページ中）で繰り返し用いられる単語は、そのページ内において重要な概念であると考えられるためである。しかしながら、ページ内に多く用いられている単語の中には、そのページを特定する性質を持たない共通あるいは汎用の単語も多く、索引語として適していないことが多い。そこで、語がどのくらい特定性を持つかをＩＤＦによって重み付けに反映させるという手法である。

【０１１４】

ＩＤＦにより、所定のデータセットの多くの文書に出現する単語の重みを小さくする効果が得られる。そのため、所定のデータセット内のページの特徴をより明確に出すことが可能となる。

【０１１５】

このＴＦ－ＩＤＦのＩＤＦに対して、本実施の形態においては、ＩＳＤＦという手法を用いている。従って、本実施の形態においては、ＴＦ－ＩＳＤＦという手法を用いて重み付けを行っていることになる。これは、ＴＦ－ＩＤＦの手法と異なり、所定の関係（この場合、兄弟関係にあるページであり、後に詳細を示すＩＣＤＦでは、共通親関係）の文書群を１つのデータセットとみなし、ＩＤＦを適用していると考えられる。

【０１１６】

すなわち、何を共通のデータセットとして見なすかが異なることになる。本実施の形態においては、兄弟関係にある文書（ページ）を１つのデータセットと見なしている。この兄弟関係にあるページとは、リンク元のページが共通という関係にある。リンク元のページが共通という関係にあるということは、そのページ間において、何らかの関係がある、何らかの類似点（共通点）があると考えられる。

【０１１７】

そのような類似点（共通点）があるページ群を１データセットとみなし、重み

付けを行う（ＩＳＤＦに基づく処理を行う）ことにより、類似したページの間の差分が、より明確になると考えられる。これにより、関連ページ検索に適した形で、各ページの特徴をより明確にすることになると考えられる。

【０１１８】

このようなことを換言すれば、どこまでを不要な特徴（雑音）として見なし、排除するかを適切に設定することにより、類似する文書に含まれる単語の重みを減じ、それらの文書（ページ）の他の特徴を浮き出させる。このように他の特徴を浮き出させることにより、類似度ではなく、関連度を求めるためのページの重み付け（特徴抽出）を行うことが可能となる。

【０１１９】

つまり、ＴＦ－ＩＤＦのＩＤＦは、あるデータセット内のページに共通で用いられる単語を不要な特徴とみなし、各ページの特徴を明確にすることで、キーワードを入力し検索結果を出力する従来の検索エンジンに適したページの特徴抽出方法として用いられてきた。しかしながら、ＴＦ－ＩＳＤＦのＩＳＤＦは、類似点がある兄弟関係のページ群をデータセットとみなし、その中で共通に用いられる単語を不要な特徴とみなすことで、関連ページ検索に適した特徴抽出手法であるといえる。

【０１２０】

このような重み付けが行われた結果が用いられて、関連度が、例えば、ＶＳＭのcosine類似度などに基づいて算出される。この関連度の算出は、上述した実施の形態においては、関連度算出部１８４により行われる。ＶＳＭについて簡便に説明するに、ＶＳＭによる手法は、出現する単語の有無や出現数を特徴量とし、検索対象データや入力文書を全単語次元数のベクトルで表現するものである。ＶＳＭでは、データ間の類似度（共通する度合い）を算出するために、ベクトル間のcosineを用いることが多い。ＶＳＭによる手法は、記事と語彙の関係、記事同士の関係、単語同士の関係をモデル化するのに有効な手法とされている。

【０１２１】

本実施の形態において、上述したような重み付けを行い、関連度を算出し、その関連度を用いて、ユーザに対して関連ページの情報を提供するため、例えば、

所定のミュージシャンのサイト内のプロフィールのページを閲覧しているときに、そのページに関連するページを検索した場合、そのミュージシャンの他のサイト内の同一プロフィールのページが検索結果としてユーザに提供されるというのではなく、そのミュージシャンの過去に参加したイベントに関する情報や、プロフィールに記載されたストーリーに関する情報、ミュージシャンが好む事柄に関する情報などの情報がユーザに提供されることになる。

【０１２２】

従って、本実施の形態によれば、ユーザが所望する関連ページをより高い精度で提供することが可能となる。

【０１２３】

一方、本実施の形態における、ページの兄弟関係、あるいは詳細を後述する共通親関係を用いたページの特徴抽出手段は、ユーザのブラウジング履歴のなかの所定のページを用いたユーザモデル生成法に適用可能である。すなわち、ユーザモデルの生成法は、ユーザが過去に参照したページ群を解析することによって生成されることが多いが、そのページの特徴抽出手段として、本実施の形態にある兄弟あるいは共通親関係のページを考慮したページの特徴抽出手段が利用できる。さらに、キーワードや自然言語を入力とした検索エンジンへ適用し、兄弟関係、あるいは共通親関係を考慮したページモデルに基づく検索エンジンの実現も可能である。

【０１２４】

上述した実施の形態においては、リンク判定部１４６（図６）は、親ページに注目して、その親ページがリンクを張っている他の子ページを判定するようにし、その結果を用いて後段の処理が行われるとしたが、子ページに注目して、その子ページにリンクを張っている他の親ページを判定するようにし、その結果を用いて後段の処理が行われるようにしても良い。

【０１２５】

すなわち、図２５を参照して説明するに、所定の子ページに注目した際、その子ページにリンクを張っている複数の親ページ（共通親のページ）が存在している場合が考えられ、それらの共通親(Co-Parent)ページの関係を、リンク関係情

報生成部１８１に相当する部分が判定し、その判定結果が用いられて、後段の処理が行われるようにしても良い。

【０１２６】

そのような判定結果を用いるようにした場合について説明する。検索サーバ４の内部構成は、基本的に、図５乃至図７に示したような構成と同様に構成することが可能である。ただし、図７に示した部分に関する構成は、図２６に示したような構成となる。図７に示した構成と、図２６に示した構成とを比較するに、図２６に示した構成は、図７のＳＤＦデータ生成部１８２とＳＤＦデータ記憶部１９２を、それぞれＣＤＦデータ生成部２５２とＣＤＦデータ記憶部２６２に置き換えた構成とされ、他の部分は、同じ構成とされている。しかしながら、各部で処理されるデータが異なり、その異なる部分について、以下に説明する。

【０１２７】

図２６に示した構成を含む検索サーバ４の動作は、図２７に示したフローチャートの処理に従って行われる。ここで、図２７に示したフローチャートを参照して、図２６に示した構成を含む検索サーバ４の動作について説明する。ステップＳ７１乃至Ｓ７６の処理は、図１１に示したフローチャートのステップＳ４１乃至Ｓ４６の処理と同様の処理であるので、その説明は省略する。

【０１２８】

ステップＳ７１乃至Ｓ７６における処理、すなわち、検索サーバ４内の構成のうち、図６に示した部分で行われる処理が行われることにより、図６に示した、保存ページ記憶部１０２、ページＩＤ記憶部１６１、単語ＩＤ記憶部１６２、基本ページモデル記憶部１６３、および、リンク情報記憶部１６４にはそれぞれ、図１４乃至図１８に示したデータが記憶される。

【０１２９】

ステップＳ７７において、リンク関係情報が、リンク関係情報生成部２５１により生成されるわけだが、その生成され、リンク関係情報記憶部２６１に記憶されるデータは、図２８に示したようなデータである。図２８に示したように、リンク関係情報記憶部２６１には、“ページＩＤ”、“Co-Parentページ ＩＤ”、および、“リンクの重み”が、それぞれ関連付けられて記憶されている。

【０１３０】

リンク関係情報生成部２５１は、各々のページＩＤに対して、Co-Parentの関係にあるページＩＤを抽出するといった処理を行うとともに、Co-Parentページ間のリンクの重みも算出する。そのCo-Parentページ間のリンクの重みの算出は、以下のようにして行われる。すなわち、Co-Parentページ間のリンクの重みWo（u,v）は、次式（１１）に基づいて算出される。

【０１３１】

$$Wo(u,v) = Wc(u,w) \times Wc(v,w) \quad \cdots (11)$$

式（１１）において、u,v,wは、Ｐをページ集合とした場合、u,v,w∈Ｐを満たす値であり、Wo（u,v）は、１≦Ｗ（u,v）を満たす値である。

【０１３２】

式（１１）において、Wo（u,v）は、所定のページuと、そのページuとCo-Parentの関係にあるCo-Parentページv間のリンクの重みであり、Wc（u,w）は、所定のページuと、そのページuと親子関係にある子ページw間のリンクの重みであり、Wc（v,w）は、所定のページvと、そのページvと親子関係にある子ページw間のリンクの重みである。

【０１３３】

このようにして、Co-Parentページ間のリンクの重みが算出され、その算出結果が、図２８に示したようなリスト形式のデータに書き込まれる。

【０１３４】

図２７のフローチャートの説明に戻り、ステップＳ７８において、ＣＤＦデータの生成が、ＣＤＦデータ生成部２５２（図２６）により行われる。ＣＤＦデータ生成部２５２は、必要に応じ、リンク関係情報記憶部２５１（図２６）と基本ページモデル記憶部１６３（図６）からデータを読み出し、その読み出したデータを用いて、図２９に示すようなリスト形式のデータを作成し、ＣＤＦデータ記憶部２６２に記憶させる。

【０１３５】

ここで、ＣＤＦとは、Co-Parent Document Frequencyの略であり、各々のページに含まれる（各ページで出現する）単語に対して、その単語が現れる共通親ペ

ージのリンクの重みを総和したデータである。

【０１３６】

図２９に示したＣＤＦデータ記憶部２６２に記憶されるデータは、“ページＩＤ”と、“ページＩＤに含まれる単語ＩＤと、その単語ＩＤを含むCo-Parentページのリンクの重みの総和”といった情報を含む。このデータは、各々のページにおいて、そのページ内で出現する単語に対して、その単語が現れるCo-Parentページのリンクの重みを総和したデータであり、リンク判定部１４６が、前述のとおりWc(p,q)＝１とリンクの重みを生成した場合には、単にその単語が現れるCo-Parentページの総数となる。

【０１３７】

ステップＳ７９において、ページモデル拡張部２５３（図２６）は、ページモデル拡張処理を実行する。ページモデル拡張処理とは、図２２に示すようなリスト形式のデータを作成し、ページモデル拡張データ記憶部２６３に記憶させる処理である。ページモデル拡張部２５３は、図２２に示すようなデータを作成するために、基本ページモデル記憶部１６３、リンク情報記憶部１６４、リンク関係情報記憶部２６１、および、ＣＤＦデータ記憶部２６２に記憶されているデータを必要に応じて読み出す。

【０１３８】

図２２に示したページモデル拡張データ記憶部２６３に記憶されているデータは、既に説明したように、“ページＩＤ”と“ベクトル”といった情報を含む。既に説明した実施の形態においては、Siblingの関係に注目したときのデータであったが、この実施の形態においては、Co-Parentの関係に注目したときのデータである。従って、そのデータの算出（“ベクトル”という情報内の“重み”という情報）に用いられる式が異なる。その異なる式に関して説明する。

【０１３９】

基本的に、Co-Parentの関係に注目し、ＩＣＤＦ（Inverse Co-Parent Document Frequency）に基づいて重みを計算した場合でも、“ベクトル”の重みに関するデータは、式（４）に基づいて算出される。ただし、式（４）に含まれるWijは、次式（１２）に基づいて算出される。

$$Wij=(1+\log(TFij))\times(1+\log(1/(1+CDFij)))\quad\cdots(12)$$

式（１２）において、TFij は、単語jのページiにおける出現回数を示し、0 ≦ TFijの値を取る。CDFij は、ページiのCo-Parentページのうち、単語jを含むページのリンクの重みの総和を示す。

【０１４０】

このような式（４）と式（１２）を用いてベクトル内の重みを算出するようにしても良いが、さらに、CDFijの効果を高めるため、式（１２）を式（１３）に置き換えても良い。

$$Wij=(1+\log(TFij))\times(1+\log(1+ACDFi/(1+CDFij)))\quad\cdots(13)$$

【０１４１】

式（１３）において、ACDFiは、ページiと全Co-Parentページ間のリンクの重みの総和を示す。

【０１４２】

さらに、TTFij, ATFijを加え、式（１２）を基に次式（１４）あるいは、式（１３）を基に次式（１５）に基づいて、重みを算出するようにしても良い。

$$Wij=(1+\log(TFij+TTFij+ATFij))\times(1+\log(1/(1+CDFij)))\quad\cdots(14)$$

$$Wij=(1+\log(TFij+TTFij+ATFij))\times(1+\log(1+ACDFi/(1+CDFij)))\quad\cdots(15)$$

式（１４）、（１５）において、TTFijは、タグ付単語jのページiにおいて出現するか否かを示し、出現しない場合０が、出現する場合1が、それぞれ設定される。あるいは出現回数(0以上)が設定されるようにしても良い。タグの種類に応じてそれぞれ重みを付けるようにしても良い。

【０１４３】

また、ATFijは、単語jのページiへのリンク元ページにおけるアンカー窓内において単語jが出現するか否かを示し、出現しない場合0が、出現する場合１が、それぞれ設定される。あるいは出現回数(0以上)が設定されるようにしても良い。タグ付単語と同様に、重み付けを行うようにしてもよい。さらに、アンカーか

らの距離に応じたウインドウ重みをつけるようにしても良い。

【０１４４】

ステップＳ８０において、関連度算出部２５４において、ページ間の関連度が算出される。関連度算出部２５４は、必要に応じ、ページモデル拡張データ記憶部２６３に記憶されているデータを読み出し、図２３に示すようなリスト形式のデータを作成し、関連度データ記憶部２６４に記憶させる。

【０１４５】

図２３に示した関連度データ記憶部２６４に記憶されるデータは、既に説明したように、“ページＩＤ”、“対象ページＩＤ”、“関連度”、および、“高関連度単語”といった情報を含む。このうち、関連度は、Co-Parentの関係に注目して処理が行われる際でも、Siblinの関係に注目して処理が行われる際と同様の式により行われる。すなわち、既に説明した式（９）に基づいて算出される。

【０１４６】

ステップＳ８１以降の処理は、図１１のステップＳ５１以降の処理と同様であるので、その説明は省略する。

【０１４７】

このように、Co-Parentの関係に注目して処理を行う場合においても、Siblingの関係に注目して処理を行う場合と同様の効果、又は、それ以上の効果を得ることが可能である。

【０１４８】

さらに、第３の実施の形態として、Siblingの関係とCo-Parentの関係の両方を考慮して処理を行うことが考えられる。そのようにした場合においても、検索サーバ４の構成は、図５乃至図７に示したような構成でよい。ただし、図７（図２６）に示した詳細な構成は、図３０に示したような構成とする。

【０１４９】

図３０に示した検索サーバ４に含まれる内部構成例について、既に説明した図７又は図２６と比較して説明する。図７に示したリンク関係情報生成部１８１または図２６に示したリンク関係情報生成部２５１は、Siblingリンク関係情報生成部３０１とCo-Parentリンク関係情報生成部３０２で構成される。またこれら

の各部で生成されたデータを記憶するために、関連ページデータ記憶部１０５には、Siblingリンク関係情報記憶部３１１とCo-Parentリンク関係情報記憶部３１２とが、それぞれ設けられている。

【０１５０】

図７に示したＳＤＦデータ生成部１８２または図２６に示したＣＤＦデータ生成部２５２は、ＳＤＦ・ＣＤＦデータ生成部３０３で構成される。また、図７に示したページモデル拡張部１８３または図２６に示したページモデル拡張部２５３は、ＩＳＤＦ・ＩＣＤＦページモデル拡張部３０４で構成される。これらの各部で生成されたデータを記憶するために、関連ページデータ記憶部１０５には、ＳＤＦ・ＣＤＦデータ記憶部３１３とＩＳＤＦ・ＩＣＤＦページモデル拡張データ記憶部３１４が、それぞれ設けられている。

【０１５１】

その他の部分に関しては、基本的に、図７（図２６）に示した構成と同様なので、その説明は省略する。

【０１５２】

図３１のフローチャートを参照して、図３０に示した構成を含む検索サーバ４の動作について説明する。ステップＳ１０１乃至Ｓ１０６の処理は、図１１に示したフローチャートのステップＳ４１乃至Ｓ４６の処理と同様の処理であるので、その説明は省略する。

【０１５３】

ステップＳ１０１乃至Ｓ１０６における処理、すなわち、検索サーバ４内の構成のうち、図６に示した部分で行われる処理が行われることにより、図６に示した、保存ページ記憶部１０２、ページＩＤ記憶部１６１、単語ＩＤ記憶部１６２、基本ページモデル記憶部１６３、および、リンク情報記憶部１６４にはそれぞれ、図１４乃至図１８に示したデータが記憶される。

【０１５４】

ステップＳ１０７において、Siblingリンク関係情報が、Siblingリンク関係情報生成部３０１（図３０）により生成されるわけだが、その生成され、Siblingリンク関係情報記憶部３１１に記憶されるデータは、図１９に示したようなデー

タである。すなわち、ステップＳ１０７における処理は、図１１のステップＳ４７の処理と同様であり、Siblingリンク関係情報生成部３０１が生成するデータは、図７に示したリンク情報関係情報生成部１８１が生成するデータと同様であるので、その詳細な説明は既に説明したので、ここではその説明を省略する。

【０１５５】

次に、ステップＳ１０８において、Co-ParentTリンク関係情報が、Co-Parentリンク関係情報生成部３０２により生成されるわけだが、その生成され、Co-Parentリンク関係情報記憶部３１２に記憶されるデータは、図２８に示したようなデータである。すなわち、ステップＳ１０８における処理は、図２７のステップＳ７７の処理と同様であり、Co-Parentリンク関係情報生成部３０２が生成するデータは、図２６に示したリンク情報関係情報生成部２５１が生成するデータと同様であるので、その詳細な説明は既に説明したので、ここではその説明を省略する。

【０１５６】

図３１のフローチャートの説明に戻り、ステップＳ１０９において、ＳＤＦ・ＣＤＦデータの生成が、ＳＤＦ・ＣＤＦデータ生成部３０３（図３０）により行われる。ＳＤＦ・ＣＤＦデータ生成部３０３は、必要に応じ、Siblingリンク関係情報記憶部３１１、Co-Parentリンク関係情報記憶部３１２、および基本ページモデル記憶部１６３（図６）からデータを読み出し、その読み出したデータを用いて、図２１と図２９に示すようなリスト形式のデータを作成し、ＳＤＦ・ＣＤＦデータ記憶部３１３に記憶させる。

【０１５７】

図２１に示したデータは、ＳＤＦ用のデータであり、図２９に示したデータは、ＣＤＦ用のデータである。ＳＤＦ用のデータは、図７のＳＤＦデータ生成部１８２が図１１のステップＳ４８の処理として行う処理と同様な処理により生成され、ＣＤＦ用のデータは、図２６のＣＤＦデータ生成部２５２が図２７のステップＳ７８の処理として行う処理と同様な処理により生成される。これらの生成については、既に説明したので、ここでは、その説明を省略する。

【０１５８】

また、図２１と図２９に示したリスト形式のデータは、それぞれ別々のリスト形式のデータとして、ＳＤＦ・ＣＤＦデータ記憶部３１３に記憶されるようにしても良いし、１つのリスト形式としてまとめられて記憶されるようにしても良い。

【０１５９】

ステップＳ１１０において、ＩＳＤＦ・ＩＣＤＦページモデル拡張部３０４（図３０）は、ＩＳＤＦ・ＩＣＤＦページモデル拡張処理を実行する。ＩＳＤＦ・ＩＣＤＦページモデル拡張処理とは、図２２に示すようなリスト形式のデータを作成し、ＩＳＤＦ・ＩＣＤＦページモデル拡張データ記憶部３１４に記憶させる処理である。

【０１６０】

ＩＳＤＦ・ＩＣＤＦページモデル拡張データ記憶部３１４に記憶されているデータは、図２２に示したようなデータであるとし、その図２２に示したデータは、既に説明したように、“ページＩＤ”と“ベクトル”といった情報を含む。図２２に示したデータについては、Siblingの関係に注目したときのデータ、または、Co-Parentの関係に注目したときのデータであるとして説明した。ここでは、その両方の関係に注目したときのデータであるため、そのデータの算出（“ベクトル”という情報内の“重み”という情報）に用いられる式が異なる。その異なる式に関して説明する。

【０１６１】

基本的に、Siblingの関係とCo-Parentの関係の両方に注目したときでも、“ベクトル”の重みに関するデータは、式（４）に基づいて算出される。式（４）に含まれるWijは、次式（１６）に基づいて算出される。

$$Wij=(1+\log(TFij))\times(1+\log(1/(1+SDFij+CDFij))) \quad \cdots (16)$$

式（１６）において、TFij は、単語jのページiにおける出現回数を示し、０≦TFijの値を取る。SDFij は、ページiのSiblingページのうち、単語jを含むページのリンクの重みの総和を示し、CDFij は、ページiのCo-Parentページのうち、単語jを含むページのリンクの重みの総和を示す。

【０１６２】

このような式（４）と式（１６）を用いてベクトル内の重みを算出するようにしても良いが、さらに、SDFijとCDFijの効果を、それぞれ高めるため、式（１６）を式（１７）に置き換えて算出するようにしても良い。

Wij = (1 + log(TFij))×(1 + log(1+ACDFi+ASDFi /(1 + ASDFij+CDFij)))　・・・（１７）

【０１６３】

式（１７）において、ASDFiは、ページiと全Siblingページ間のリンクの重みの総和を、ACDFiは、ページiと全Co-Parentページ間のリンクの重みの総和を示す。

【０１６４】

さらに、TTFij，ATFijを加え、式（１６）を基に次式（１８）あるいは、式（１７）を基に次式（１９）に基づいて、重みを算出するようにしても良い。

Wij = (1 + log(TFij＋TTFij+ATFij))×(1 + log(1 /(1 + SDFij+CDFij))) ・・・（１８）

Wij = (1 + log(TFij＋TTFij+ATFij))×(1 + log(1+ ASDFi+ACDFi /(1 + SDFij+CDFij))) ・・（１９）

式（１８）または式（１９）において、TTFijは、タグ付単語jのページiにおいて出現するか否かを示し、出現しない場合０が、出現する場合1が、それぞれ設定される。あるいは出現回数(0以上)が設定されるようにしても良い。タグの種類に応じてそれぞれ重みを付けるようにしても良い。

【０１６５】

また、ATFijは、単語jのページiへのリンク元ページにおけるアンカー窓内において単語jが出現するか否かを示し、出現しない場合0が、出現する場合１が、それぞれ設定される。あるいは出現回数(0以上)が設定されるようにしても良い。タグ付単語と同様に、重み付けを行うようにしてもよい。さらに、アンカーからの距離に応じたウインドウ重みをつけるようにしても良い。

【０１６６】

図３１のフローチャートの説明に戻り、ステップＳ１１１において、関連度算

出部３０５において、ページ間の関連度が算出される。関連度算出部３０５は、必要に応じ、ISDF・ICDFページモデル拡張データ記憶部３１４に記憶されているデータを読み出し、図２３に示すようなリスト形式のデータを作成し、関連度データ記憶部３１５に記憶させる。

【０１６７】

図２３に示した関連度データ記憶部２６４に記憶されるデータは、既に説明したように、“ページＩＤ”、“対象ページＩＤ”、“関連度”、および、“高関連度単語”といった情報を含む。このうち、関連度は、Co-Parentの関係に注目して処理が行われる際でも、Siblingの関係に注目して処理が行われる際でも、または、SiblingとCo-Parentの両方の関係に注目して処理が行われる際でも、同様の式により行われる。すなわち、既に説明した式（９）に基づいて算出される。

【０１６８】

ステップＳ１１２以降の処理は、図１１のステップＳ５１以降の処理と同様であるので、その説明は省略する。

【０１６９】

このように、Siblinの関係とCo-Parentの関係の両方に注目して処理を行う場合においても、Co-Parentの関係に注目して処理を行うときや、Siblingの関係に注目して処理を行うときと同様の効果、またはそれ以上の効果を得ることが可能である。

【０１７０】

上述した実施の形態においては、ユーザに関連ページの情報を提供する際の処理について説明したが、その関連ページの情報に、広告などの情報を含めるようにしても良い。そのような広告などの情報も提供するようにした場合、検索サーバ４の構成は、図３２に示したようになる。図３２に示した検索サーバ４の構成は、図５に示した検索サーバ４の構成に、特殊設定管理用記憶部３３１を追加した構成とされている。

【０１７１】

この特殊設定管理用記憶部３３１には、図３３、図３４にそれぞれ示す記憶部

が設けられている。図３３に示した特殊設定用管理データ記憶部３４１には、“タイトル”、“リンク先ＵＲＬ”、“説明”、“単語”、“ＵＲＬパターン”、および、“オーナＩＤ”といった情報が含まれている。図３４に示した特殊設定管理者データ記憶部３４２には、“オーナＩＤ”、“名前”、“所属”、“e-mail”、“Account”、および、“Password”といった情報が含まれている。

【０１７２】

このような特殊設定管理用記憶部３３１が、検索サーバ４に設けられた場合、例えば、図１１に示したフローチャートにおいて、関連ページ一覧生成という処理の内の１処理として、この特殊設定管理用記憶部３３１に記憶されている情報を提供するための処理が実行される。具体的には、関連ページの一覧表のデータが、作成された後に、特殊設定管理用記憶部３３１が参照され、その関連ページに関連すると判断されるＵＲＬなどの情報が、特殊設定用管理データ記憶部３４１から抽出され、一覧表のデータに含まれる。

【０１７３】

提供されたデータがユーザ側の端末３で再生されると、その画面には、関連ページの一覧と、その関連ページに関わりのある情報（広告）が表示されている。

【０１７４】

特殊設定用管理データ記憶部３４１に記憶されているデータは、管理者により削除、追加、訂正などの処理が行えるようになっており、その管理者を管理するためのデータが、特殊設定管理者データ記憶部３４２に記憶されている。この特殊設定管理者データ記憶部３４２に記憶されている管理者のみが、特殊設定用管理データ記憶部３４１のデータを操作することが可能とするために、パスワード（Password）などが設定されるようになっている。

【０１７５】

このように、関連ページの一覧表に、広告も含めるようにした場合、その広告を掲載する会社から、その掲載料金を徴収することが可能となる。また、上述した実施の形態においては説明しなかったが、例えば、検索サーバ４の収集サイトリスト記憶部１０１に記憶されるサイトを管理する管理者から、料金を徴収するようにしても良い。

【０１７６】

これは、検索サーバ４により、ユーザに関連ページであるとしてユーザに提供されることにより、そのサイトへのアクセスの増加を期待することができ、そのために、検索サーバ４自体に登録してもらいたいというサイトの管理者から登録料として料金を徴収することができる。

【０１７７】

このような課金制度を、必要に応じて設けることも可能である。

【０１７８】

上述した一連の処理は、それぞれの機能を有するハードウェアにより実行させることもできるが、ソフトウェアにより実行させることもできる。一連の処理をソフトウェアにより実行させる場合には、そのソフトウェアを構成するプログラムが専用のハードウェアに組み込まれているコンピュータ、または、各種のプログラムをインストールすることで、各種の機能を実行することが可能な、例えば汎用のパーソナルコンピュータなどに、記録媒体からインストールされる。

【０１７９】

記録媒体は、図２に示すように、WWWサーバ２を構成するパーソナルコンピュータとは別に、ユーザにプログラムを提供するために配布される、プログラムが記録されている磁気ディスク３１（フレキシブルディスクを含む）、光ディスク３２（CD-ROM（Compact Disc-Read Only Memory），DVD（Digital Versatile Disc）を含む）、光磁気ディスク３３（MD（Mini-Disc）（登録商標）を含む）、若しくは半導体メモリ３４などよりなるパッケージメディアにより構成されるだけでなく、コンピュータに予め組み込まれた状態でユーザに提供される、プログラムが記憶されているＲＯＭ１２や記憶部１８が含まれるハードディスクなどで構成される。

【０１８０】

なお、本明細書において、媒体により提供されるプログラムを記述するステップは、記載された順序に従って、時系列的に行われる処理は勿論、必ずしも時系列的に処理されなくとも、並列的あるいは個別に実行される処理をも含むものである。

【０１８１】

また、本明細書において、システムとは、複数の装置により構成される装置全体を表すものである。

【０１８２】

【発明の効果】

本発明の情報処理装置および方法、記録媒体、並びにプログラムによれば、インターネット上に開設されているサイトの検索を行うことが可能である。

【０１８３】

また、本発明の情報処理装置および方法、記録媒体、並びにプログラムによれば、よりユーザの所望としているサイトを検索し、その情報を提供することが可能である。

【図面の簡単な説明】

【図１】

本発明を適用した情報処理システムの一実施の形態の構成を示す図である。

【図２】

WWWサーバの内部構成例を示す図である。

【図３】

端末３の内部構成例を示す図である。

【図４】

検索サーバの内部構成例を示す図である。

【図５】

検索サーバの内部構成例を示す図である。

【図６】

検索サーバの詳細な内部構成例を示す図である。

【図７】

検索サーバの詳細な内部構成例を示す図である。

【図８】

リンク関係について説明するための図である。

【図９】

端末と検索サーバとの間で行われる処理について説明するフローチャートである。

【図１０】

端末側のディスプレイ上に表示される画面の一例を示す図である。

【図１１】

検索サーバの動作について説明するためのフローチャートである。

【図１２】

収集サイトリスト記憶部に記憶されるデータを説明するための図である。

【図１３】

保存ページ記憶部に記憶されるサイトのデータを説明するための図である。

【図１４】

ページＩＤ記憶部に記憶されるデータを説明するための図である。

【図１５】

単語ＩＤ記憶部に記憶されるデータを説明するための図である。

【図１６】

単語ＩＤ記憶部に記憶されるデータを説明するための図である。

【図１７】

基本ページモデル記憶部に記憶されるデータを説明するための図である。

【図１８】

リンク情報記憶部に記憶されるデータを説明するための図である。

【図１９】

リンク関係情報記憶部に記憶されるデータを説明するための図である。

【図２０】

重みの算出について説明するための図である。

【図２１】

ＳＤＦデータ記憶部に記憶されているデータを説明するための図である。

【図２２】

ページモデル拡張データ記憶部に記憶されるデータを説明するための図である。

【図２３】

関連度データ記憶部に記憶されるデータを説明するための図である。

【図２４】

関連ページ間の特徴の抽出について説明するための図である。

【図２５】

リンク関係について説明するための図である。

【図２６】

検索サーバの詳細な他の内部構成例を示す図である。

【図２７】

図２６に示した構成を有する検索サーバの動作について説明するフローチャートである。

【図２８】

リンク関係情報記憶部に記憶されるデータを説明するための図である。

【図２９】

ＣＤＦデータ記憶部２６２に記憶されるデータを説明するための図である。

【図３０】

検索サーバの詳細な他の内部構成例を示す図である。

【図３１】

図３０に示した構成を有する検索サーバの動作について説明するフローチャートである。

【図３２】

検索サーバの他の内部構成例を示す図である。

【図３３】

特殊設定用管理データ記憶部に記憶されるデータを説明する図である。

【図３４】

特殊設定管理者データ記憶部に記憶されるデータを説明する図である。

【符号の説明】

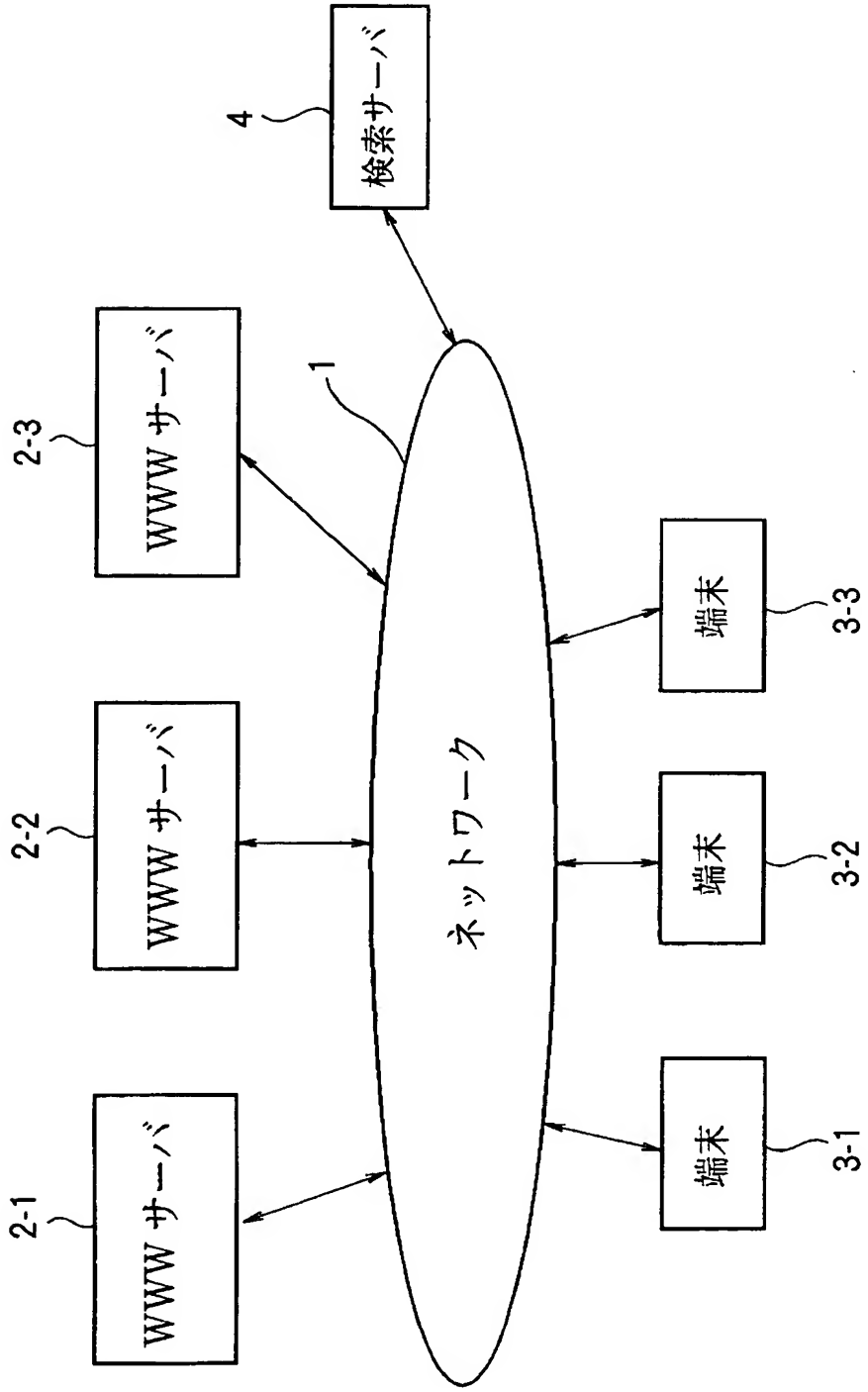
１　ネットワーク，　２　ＷＷＷサーバ，　３　端末，　４　検索サーバ，１０１　収集サイトリスト記憶部，　１０２　保存ページ記憶部，　１０３　ペ

ージデータ記憶部，　１０４　サイトページデータ記憶部，　１０５　関連ページデータ記憶部，　１１１　サイトページ記憶部，　１１２　関連ページデータ処理部

【書類名】　図面

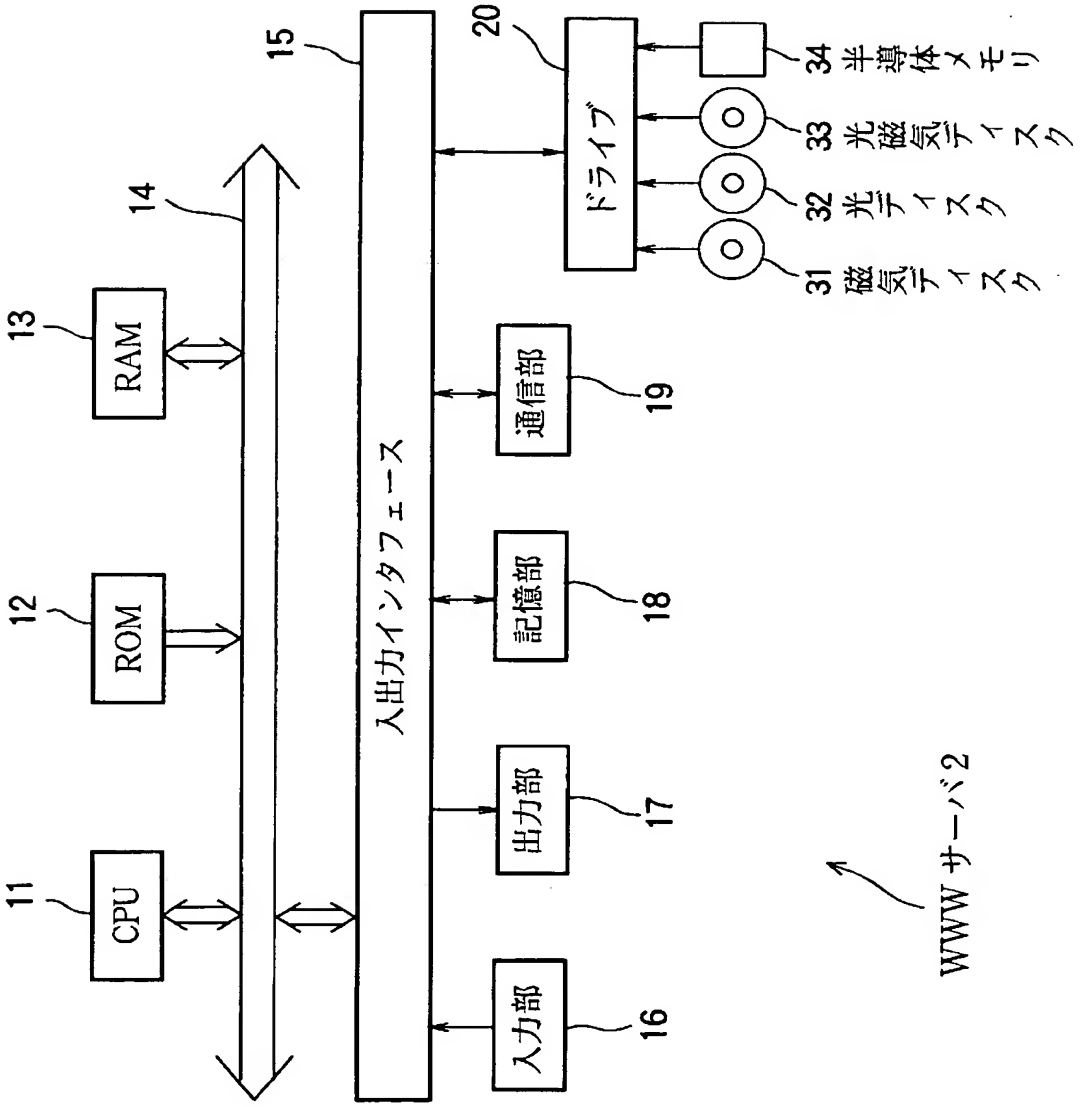
【図１】

図１



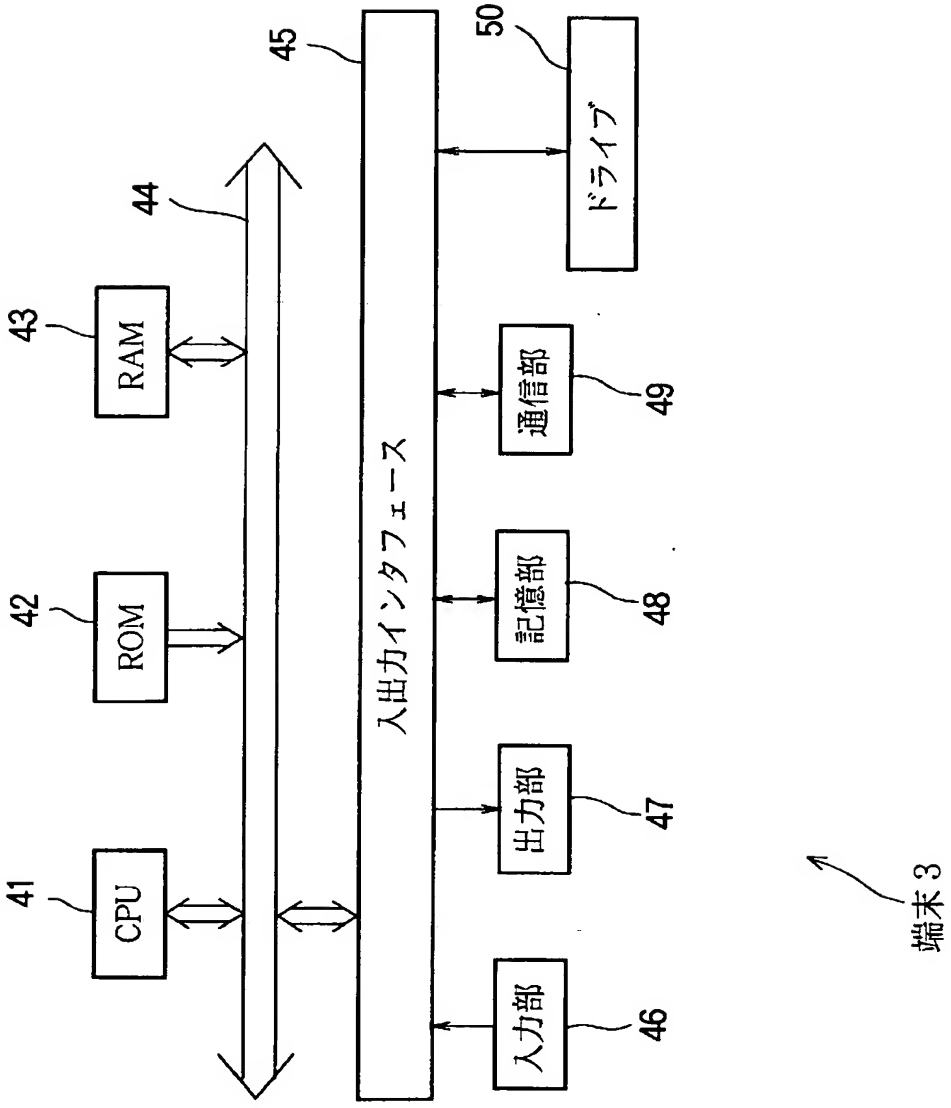
【図２】

図２


WWW サーバ2
CPU
11
ROM
12
RAM
13
14
15
入出力インタフェース
入力部
16
出力部
17
記憶部
18
通信部
19
20
ドライブ
31 磁気ディスク
32 光ディスク
33 光磁気ディスク
34 半導体メモリ

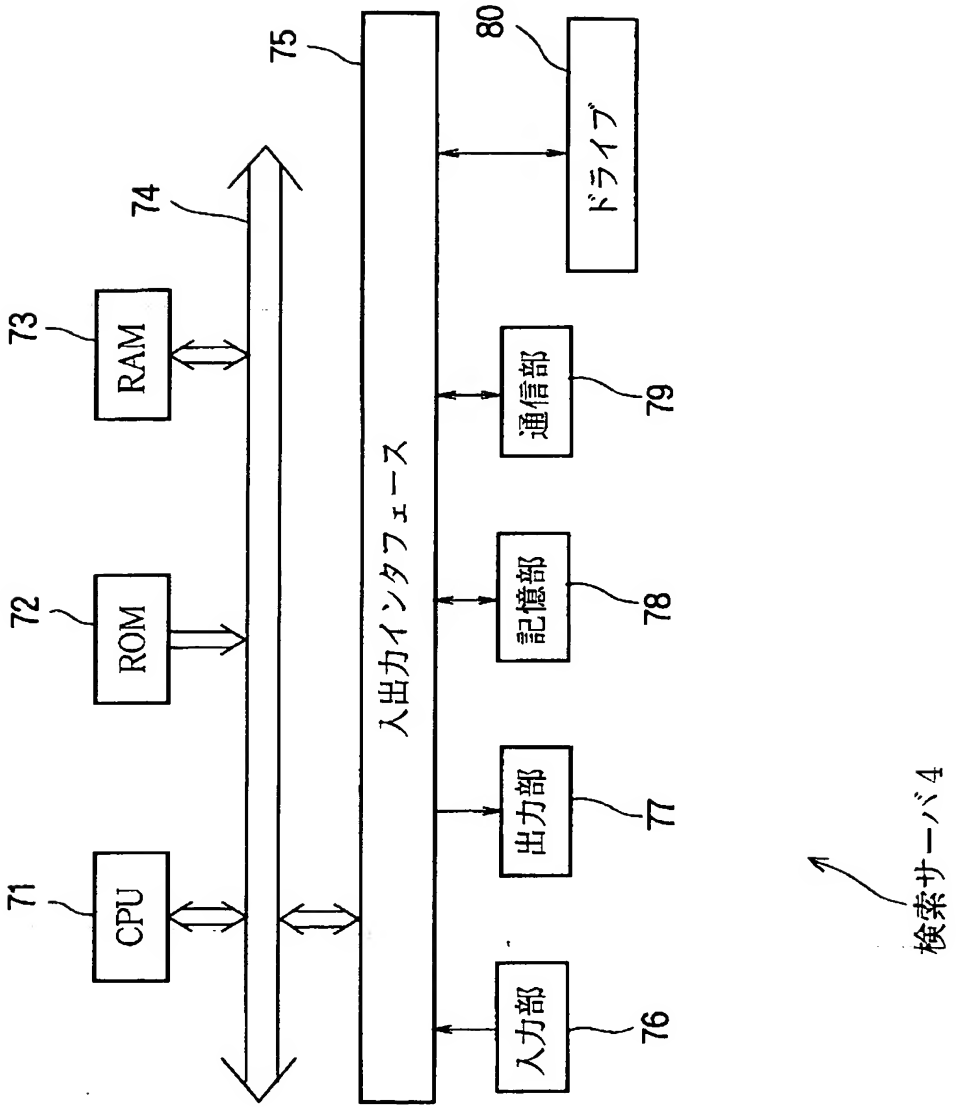
【図３】

図３


41
CPU
42
ROM
43
RAM
44
45
入出力インタフェース
46
入力部
47
出力部
48
記憶部
49
通信部
50
ドライブ
端末３

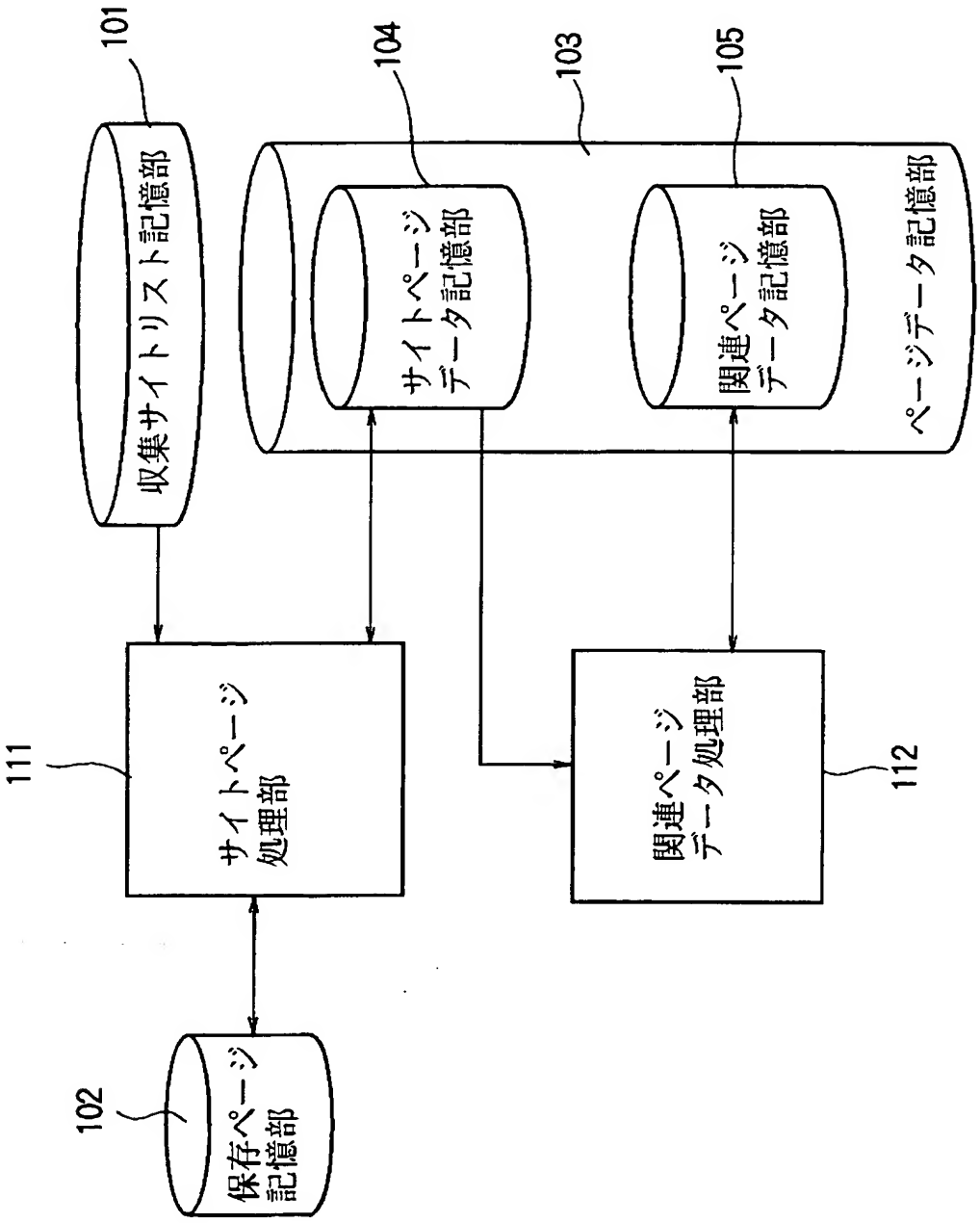
【図４】

図４


検索サーバ４
CPU
71
ROM
72
RAM
73
74
75
入出力インタフェース
入力部
76
出力部
77
記憶部
78
通信部
79
ドライブ
80

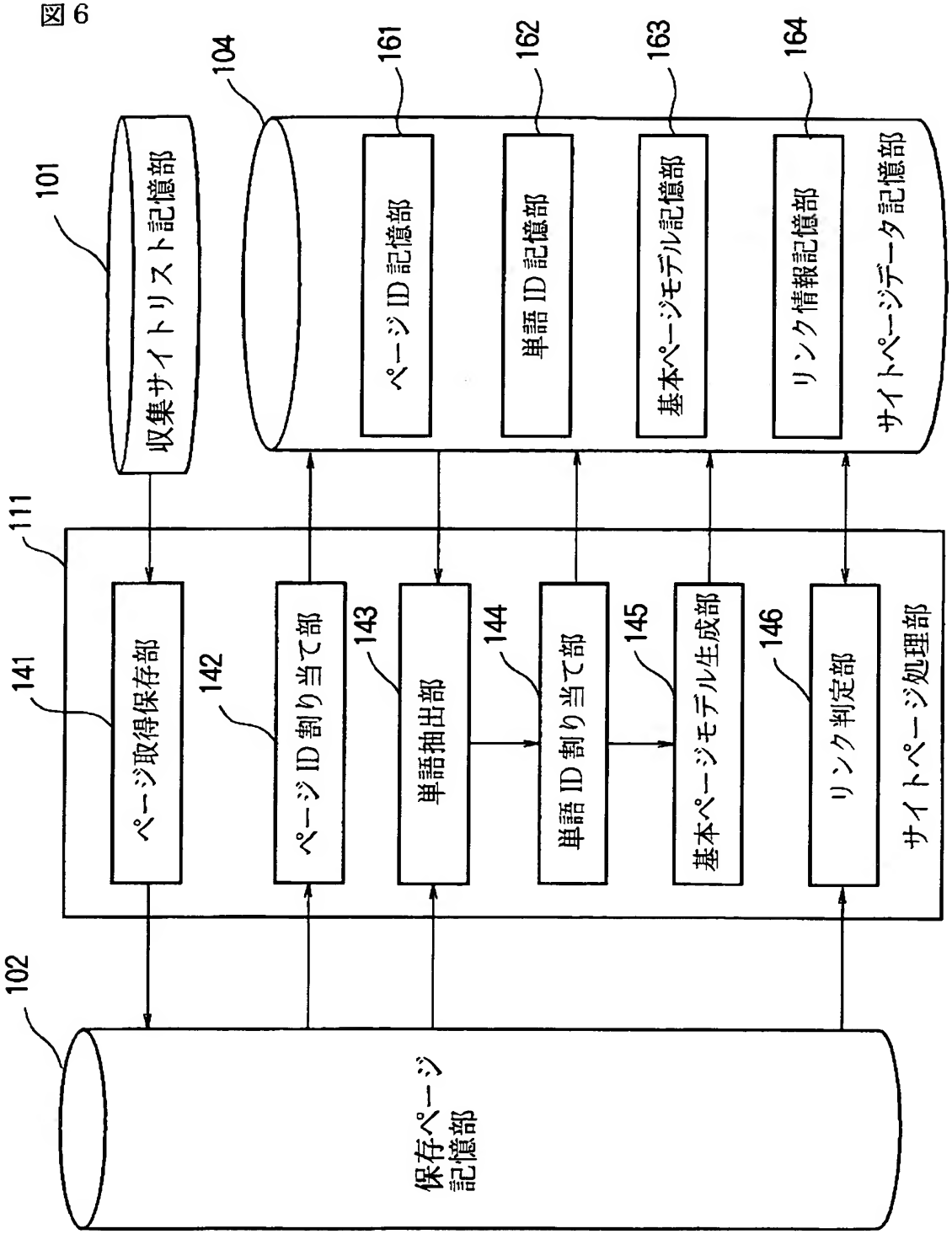
【図５】

図５


101
収集サイトリスト記憶部
104
サイトページデータ記憶部
103
105
関連ページデータ記憶部
ページデータ記憶部
111
サイトページ処理部
関連ページデータ処理部
112
102
保存ページ記憶部

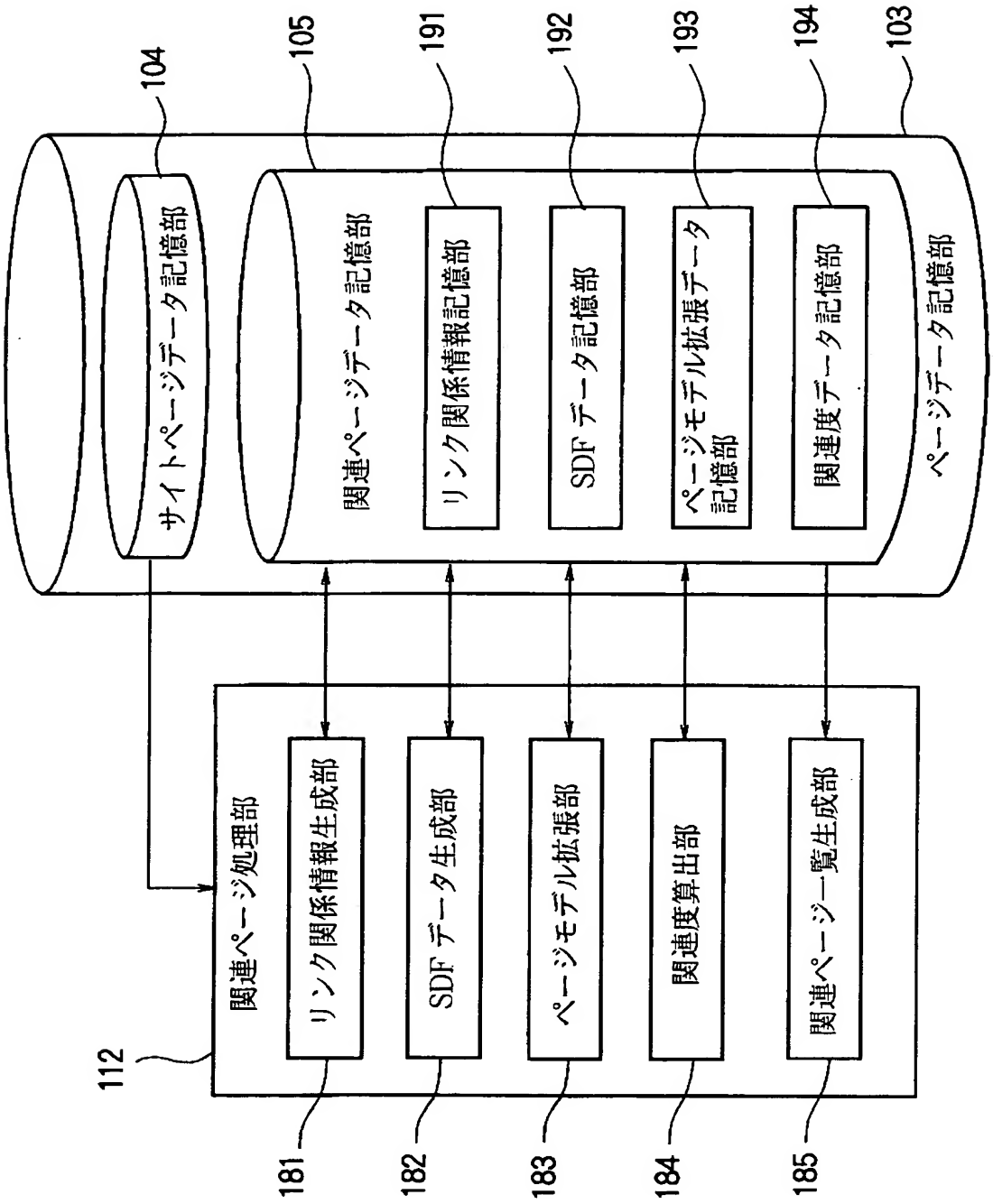
【図６】

図６


101
収集サイトリスト記憶部
102
保存ページ記憶部
104
サイトページデータ記憶部
161
ページ ID 記憶部
162
単語 ID 記憶部
163
基本ページモデル記憶部
164
リンク情報記憶部
111
サイトページ処理部
141
ページ取得保存部
142
ページ ID 割り当て部
143
単語抽出部
144
単語 ID 割り当て部
145
基本ページモデル生成部
146
リンク判定部

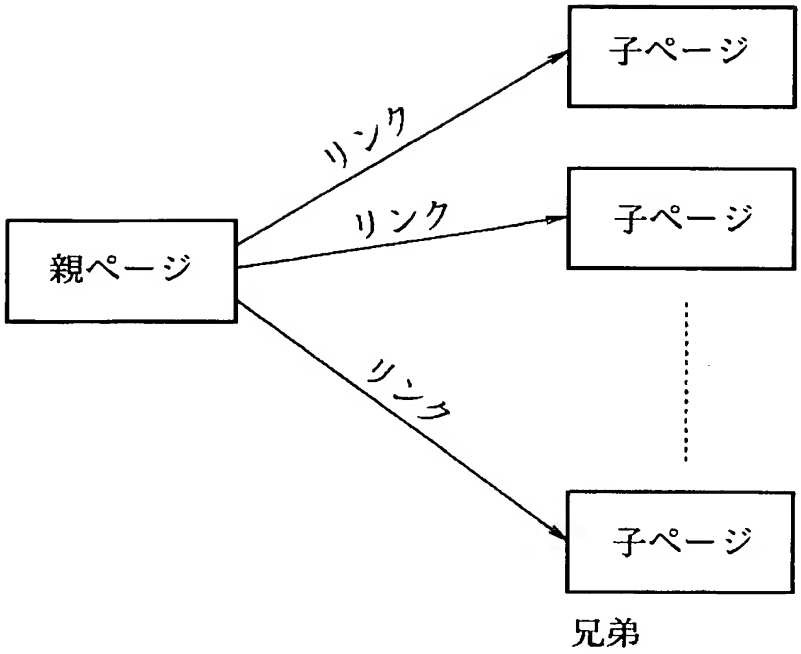
【図７】

図７


104
105
191
192
193
194
103
サイトページデータ記憶部
関連ページデータ記憶部
リンク関係情報記憶部
SDF データ記憶部
ページモデル拡張データ記憶部
関連度データ記憶部
ページデータ記憶部
112
関連ページ処理部
リンク関係情報生成部
SDF データ生成部
ページモデル拡張部
関連度算出部
関連ページ一覧生成部
181
182
183
184
185

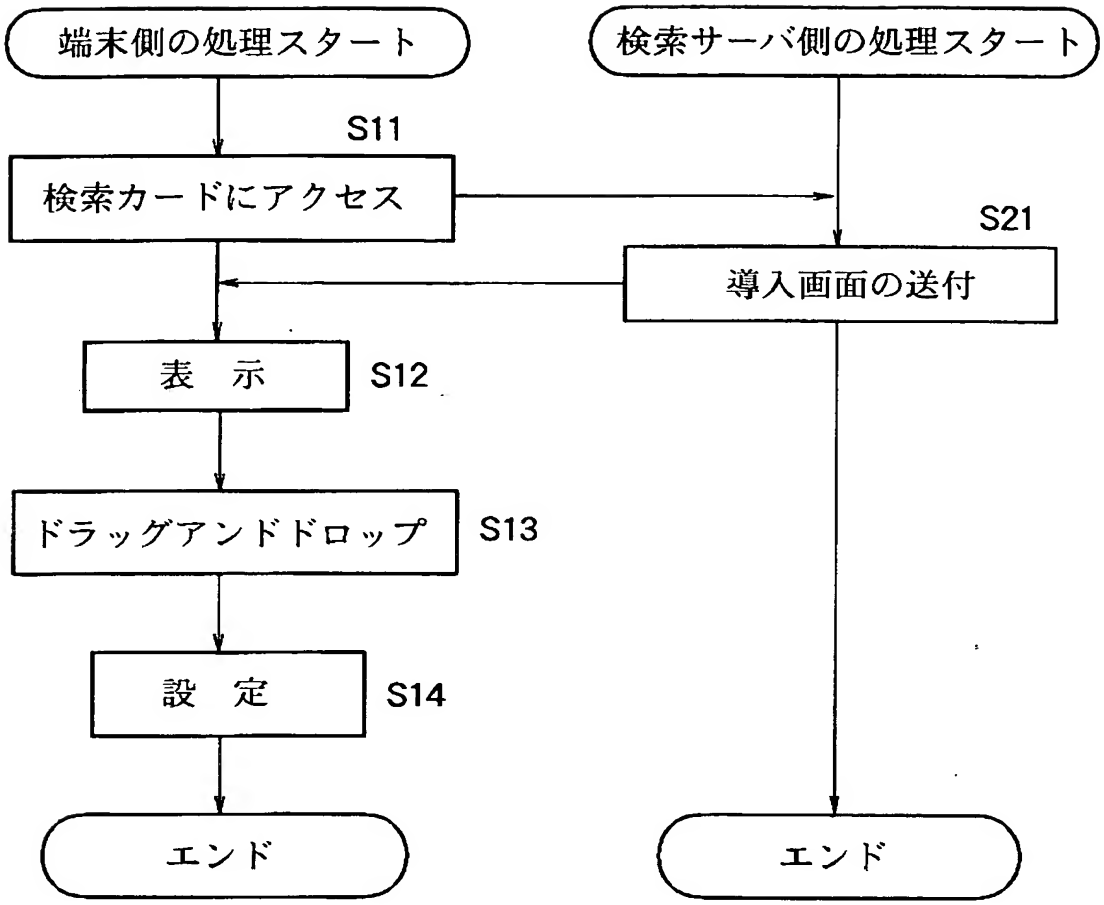
【図８】

図８



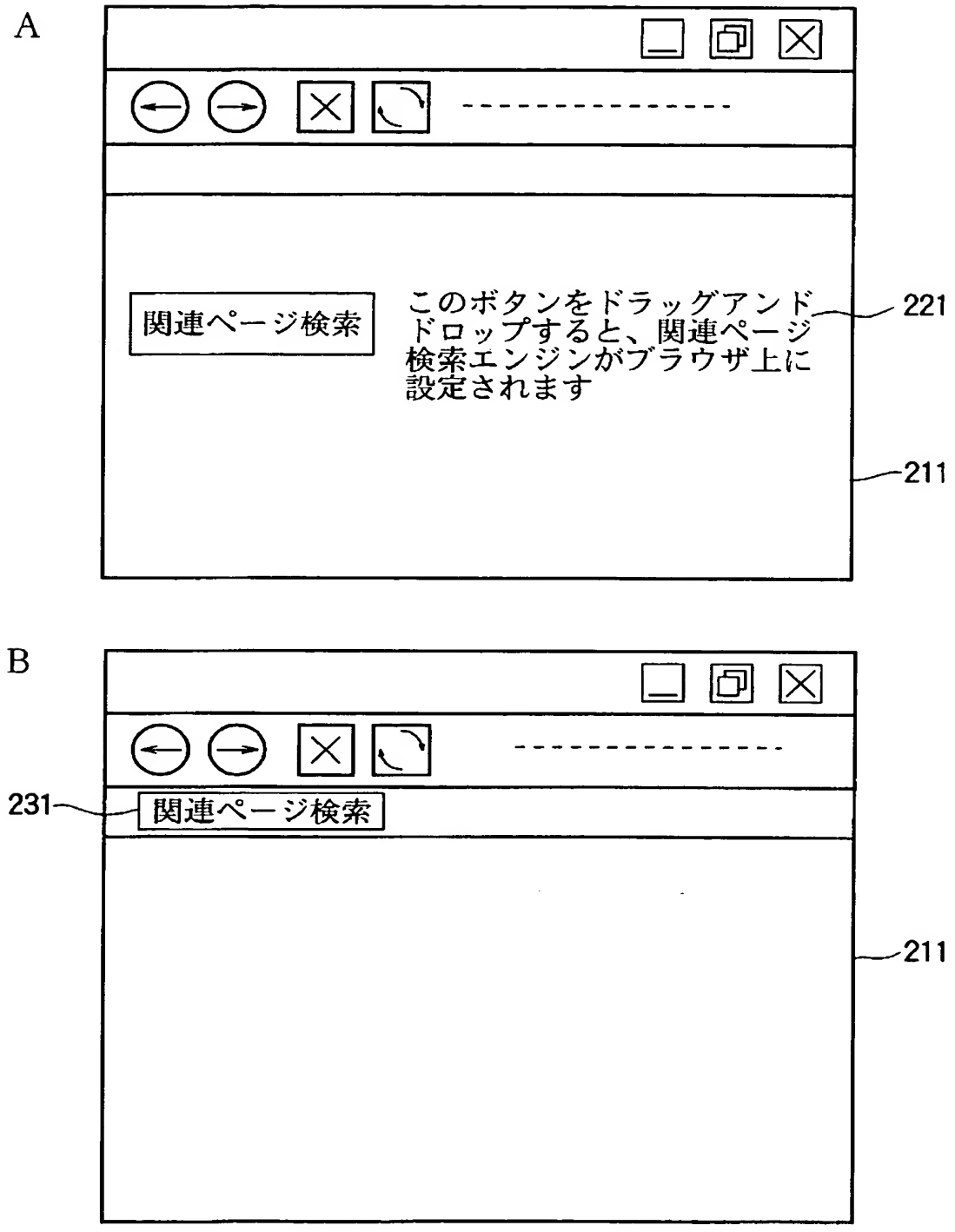
【図９】

図９


端末側の処理スタート
検索サーバ側の処理スタート
S11
検索カードにアクセス
S21
導入画面の送付
表　示
S12
ドラッグアンドドロップ
S13
設　定
S14
エンド
エンド

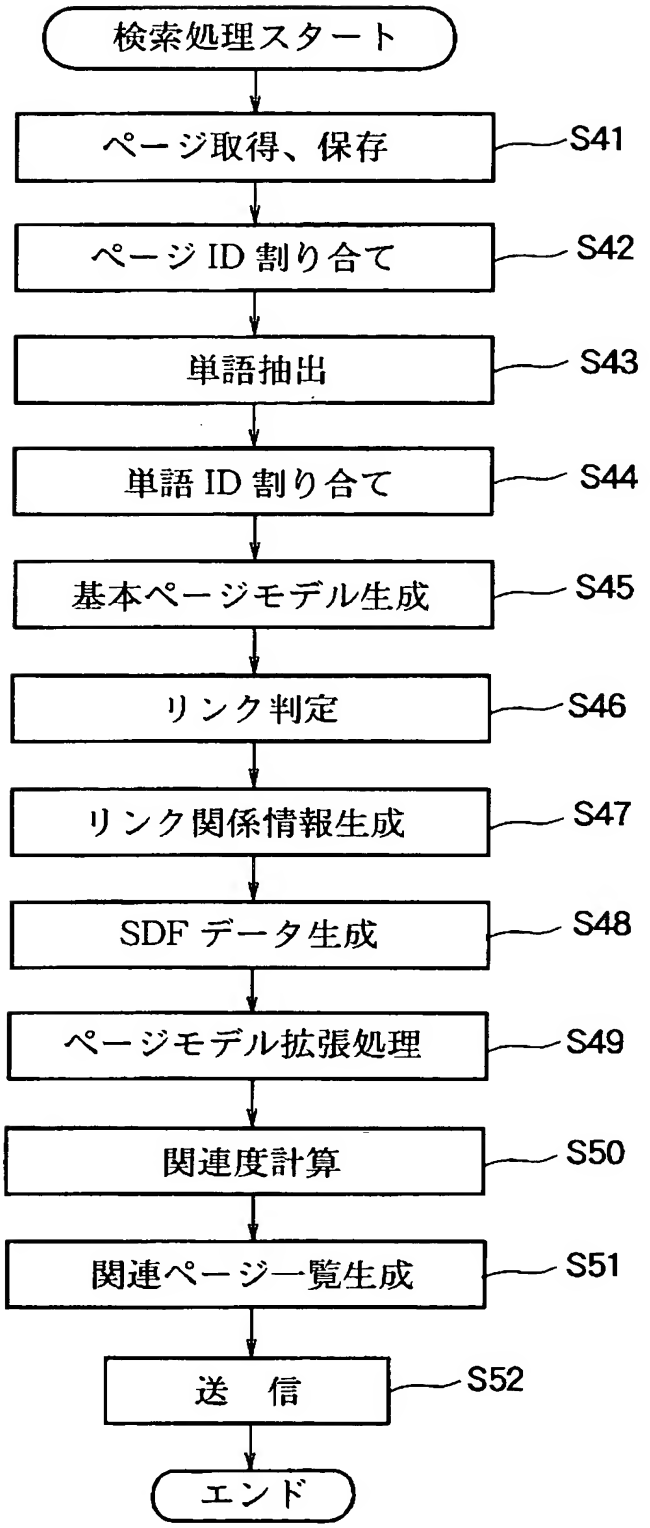
【図１０】

図 10


A
関連ページ検索
このボタンをドラッグアンドドロップすると、関連ページ検索エンジンがブラウザ上に設定されます
221
211
B
231
関連ページ検索
211

【図１１】

図11


検索処理スタート
ページ取得、保存
S41
ページID割り合て
S42
単語抽出
S43
単語ID割り合て
S44
基本ページモデル生成
S45
リンク判定
S46
リンク関係情報生成
S47
SDFデータ生成
S48
ページモデル拡張処理
S49
関連度計算
S50
関連ページ一覧生成
S51
送　信
S52
エンド

【図１２】

図 12

収集開始 URL (Key)	含む ディレクトリ	排他 ディレクトリ	含む ドメイン	排他 ドメイン
www.ssss.co.jp／index.html				
www.vaaa.sss.co.jp／index.html				
www.spe.co.jp／index.html				

101

【図１３】

図 13

サイトID(Key)	サイト名	総ページ数
0001	www.ssss.co.jp	12345
0002	www.ssss.music.co.jp	4321

102

【図１４】

図 14

ページID(Key)	サイトID	ページURL	タイトル	サマリー	ページ保存場所	最終更新日
0001000001	0001	www.ssss.co.jp／index.html	Ssss	Ssss Corp.	A／index.html	2002-07-10
0002000001	0002	www.spe.co.jp／index.html	SPE	SPE Japan	B／index.html	2002-07-10

161

【図１５】

図15

単語ID(Key)	単語
000001	映画
000002	MUSIC

162

【図１６】

図 16

単語 ID (Key1)	サイト ID (Key2)	そのサイト内で当該単語を含むページ数	そのサイト内で当該単語を含むページ ID
000001	001	30	0001000001,0001000002,.....
000001	002	15	0002000001,0002000004,.....

162

【図１７】

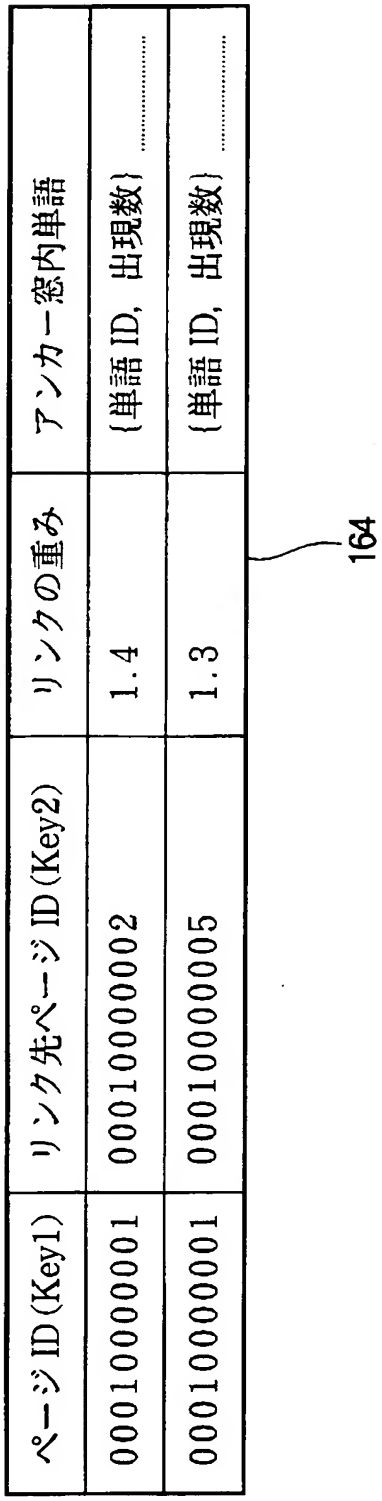
図 17

ページ ID (Key)	出現単語	Title	Keywords	description
0001000001	{単語 ID, 出現数}	{単語 ID, 出現数}	{単語 ID, 出現数}	{単語 ID, 出現数}
0001000002	{単語 ID, 出現数}	{単語 ID, 出現数}	{単語 ID, 出現数}	{単語 ID, 出現数}

163

【図１８】

図 18



ページID(Key1)	リンク先ページID(Key2)	リンクの重み	アンカー窓内単語
0001000001	0001000002	1.4	{単語ID，出現数} .....
0001000001	0001000005	1.3	{単語ID，出現数} .....

164

【図１９】

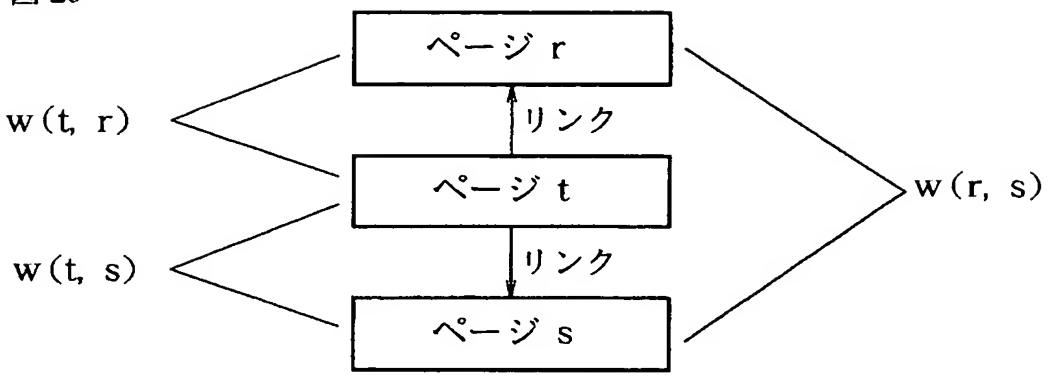
図 19

ページ ID (Key1)	Sibling ページ ID (Key2)	リンクの重み
000100000005	000100000007	1.5
000100000005	000200000002	1.3

191

【図２０】

図 20



【図２１】

図 21

ページ ID (Key)	ページ ID に含まれる単語 ID と、その単語 ID を含む Sibling ページのリンクの重みの総和
0001000001	{単語 ID, リンクの重みの総和}, {単語 ID, リンクの重みの総和}, ...
0001000002	{単語 ID, リンクの重みの総和}, {単語 ID, リンクの重みの総和}, ...

192

【図２２】

図 22

ページ ID (Key)	ベクトル
0001000001	{単語 ID,　重み}, ...
0001000002	{単語 ID,　重み}, ...

193

【図２３】

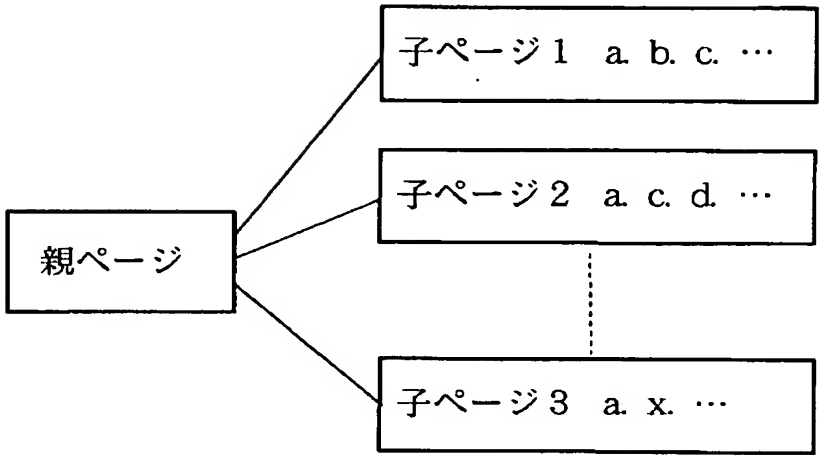
図 23

ページ ID (Key1)	対象ページ ID (Key2)	関連度	高関連度単語
0001000001	0001000002	0.10	{単語 ID,　単語関連度} .....
0001000001	0001000003	0.40	{単語 ID,　単語関連度} .....
0001000001	0001000004	0.12	{単語 ID,　単語関連度} .....

194

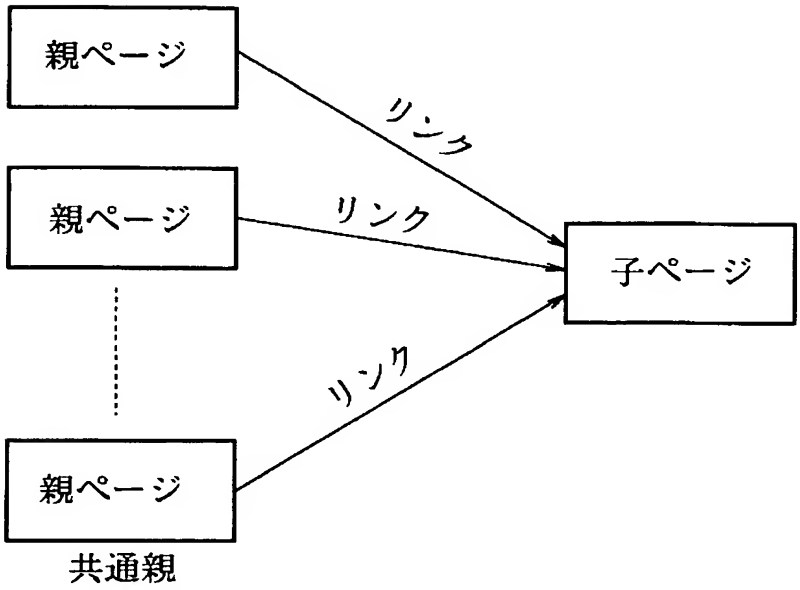
【図２４】

図 24



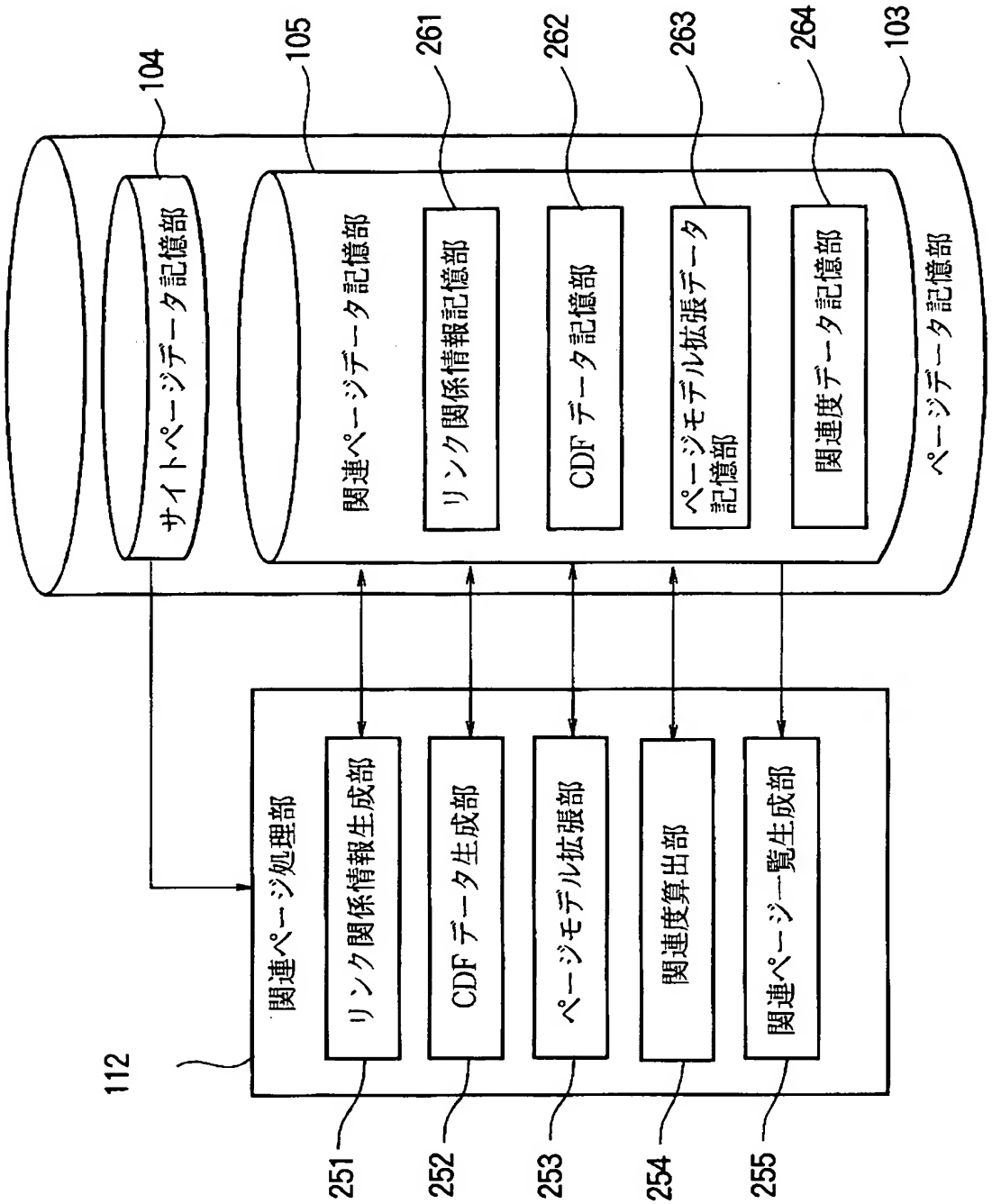
【図２５】

図 25



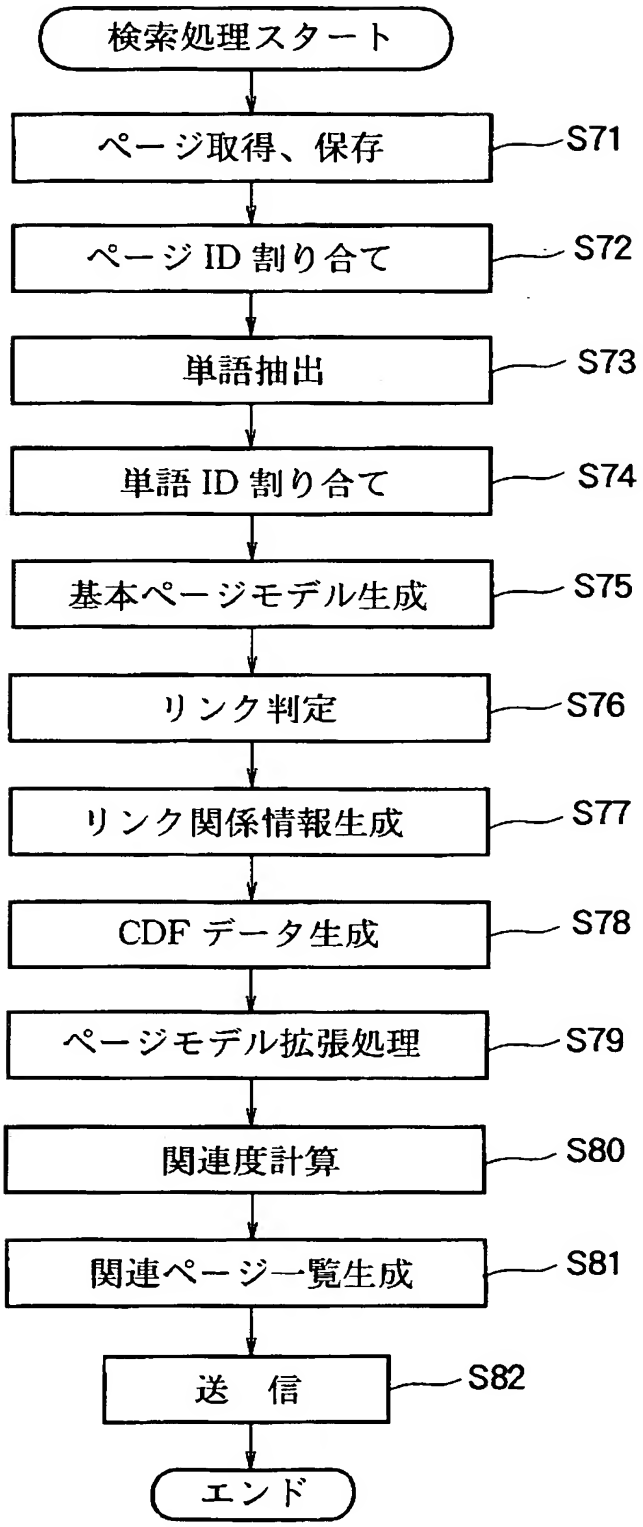
【図２６】

図 26


104
105
261
262
263
264
103
サイトページデータ記憶部
関連ページデータ記憶部
リンク関係情報記憶部
CDF データ記憶部
ページモデル拡張データ記憶部
関連度データ記憶部
ページデータ記憶部
112
関連ページ処理部
リンク関係情報生成部
CDF データ生成部
ページモデル拡張部
関連度算出部
関連ページ一覧生成部
251
252
253
254
255

【図２７】

図 27


検索処理スタート
ページ取得、保存
S71
ページ ID 割り合て
S72
単語抽出
S73
単語 ID 割り合て
S74
基本ページモデル生成
S75
リンク判定
S76
リンク関係情報生成
S77
CDF データ生成
S78
ページモデル拡張処理
S79
関連度計算
S80
関連ページ一覧生成
S81
送　信
S82
エンド

【図２８】

図 28

ページ ID (Key1)	Co-Parent ページ ID(Key2)	リンクの重み
00010000007	00010000006	1.2
00010000008	00010000006	1.4

261

【図２９】

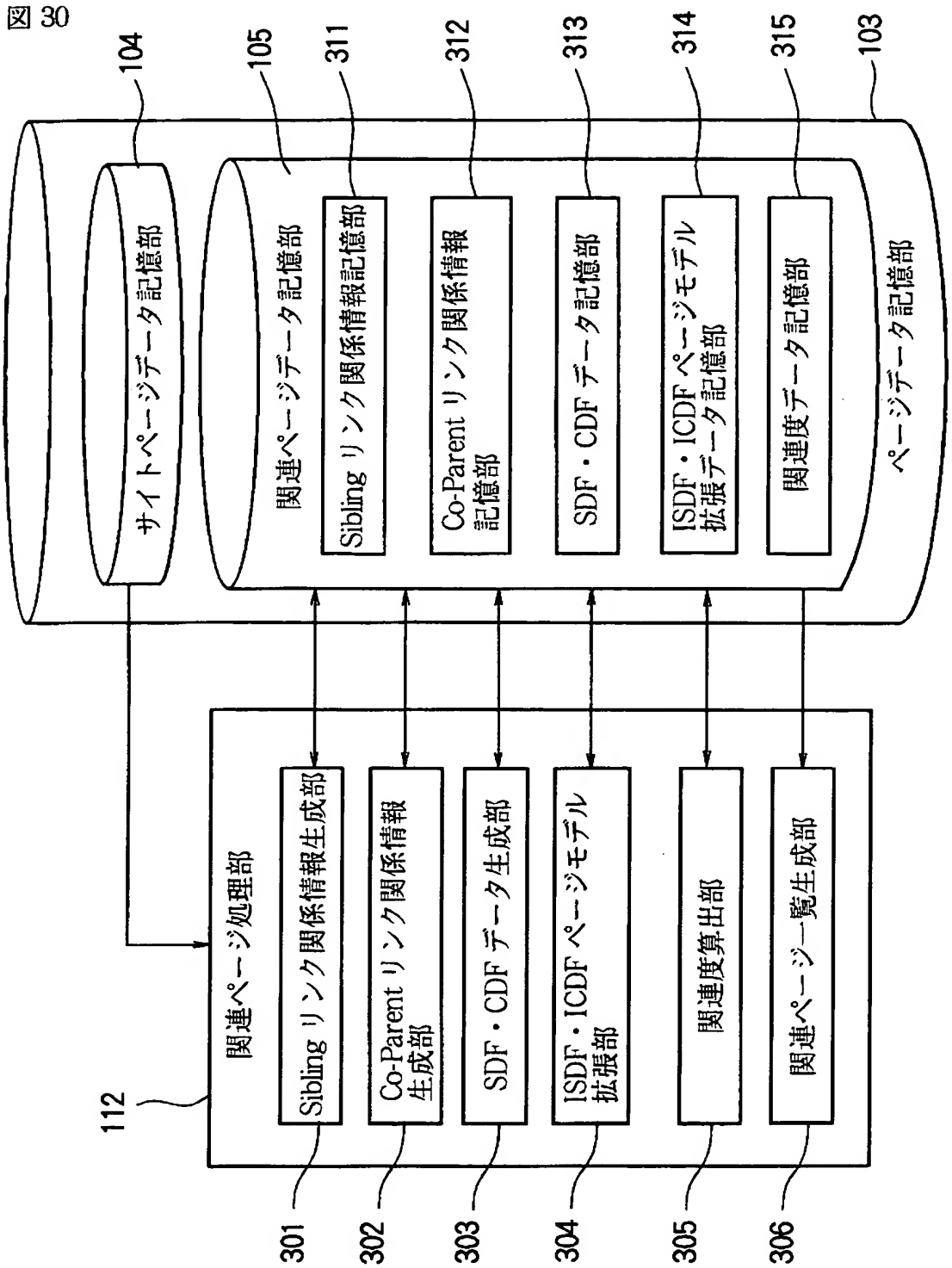
図 29

ページ ID (Key)	ページ ID に含まれる単語 ID と、その単語 ID を含む Co-Parent ページのリンクの重みの総和
00010000007	{単語 ID, リンクの重みの総和},{単語 ID, リンクの重みの総和}.....
00010000008	{単語 ID, リンクの重みの総和},{単語 ID, リンクの重みの総和}.....

262

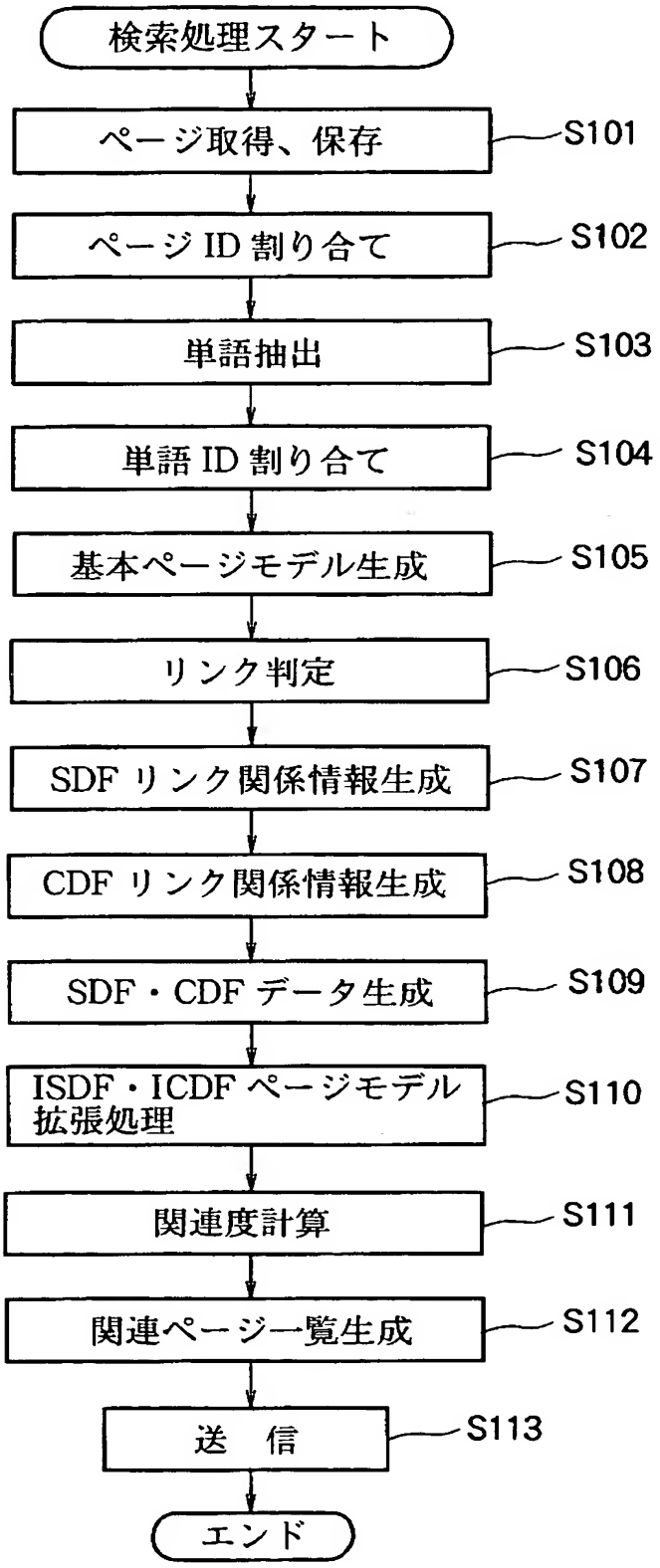
【図３０】

図 30


104
105
311
312
313
314
315
103
サイトページデータ記憶部
関連ページデータ記憶部
Sibling リンク関係情報記憶部
Co-Parent リンク関係情報記憶部
SDF・CDF データ記憶部
ISDF・ICDF ページモデル拡張データ記憶部
関連度データ記憶部
ページデータ記憶部
112
関連ページ処理部
Sibling リンク関係情報生成部
Co-Parent リンク関係情報生成部
SDF・CDF データ生成部
ISDF・ICDF ページモデル拡張部
関連度算出部
関連ページ一覧生成部
301
302
303
304
305
306

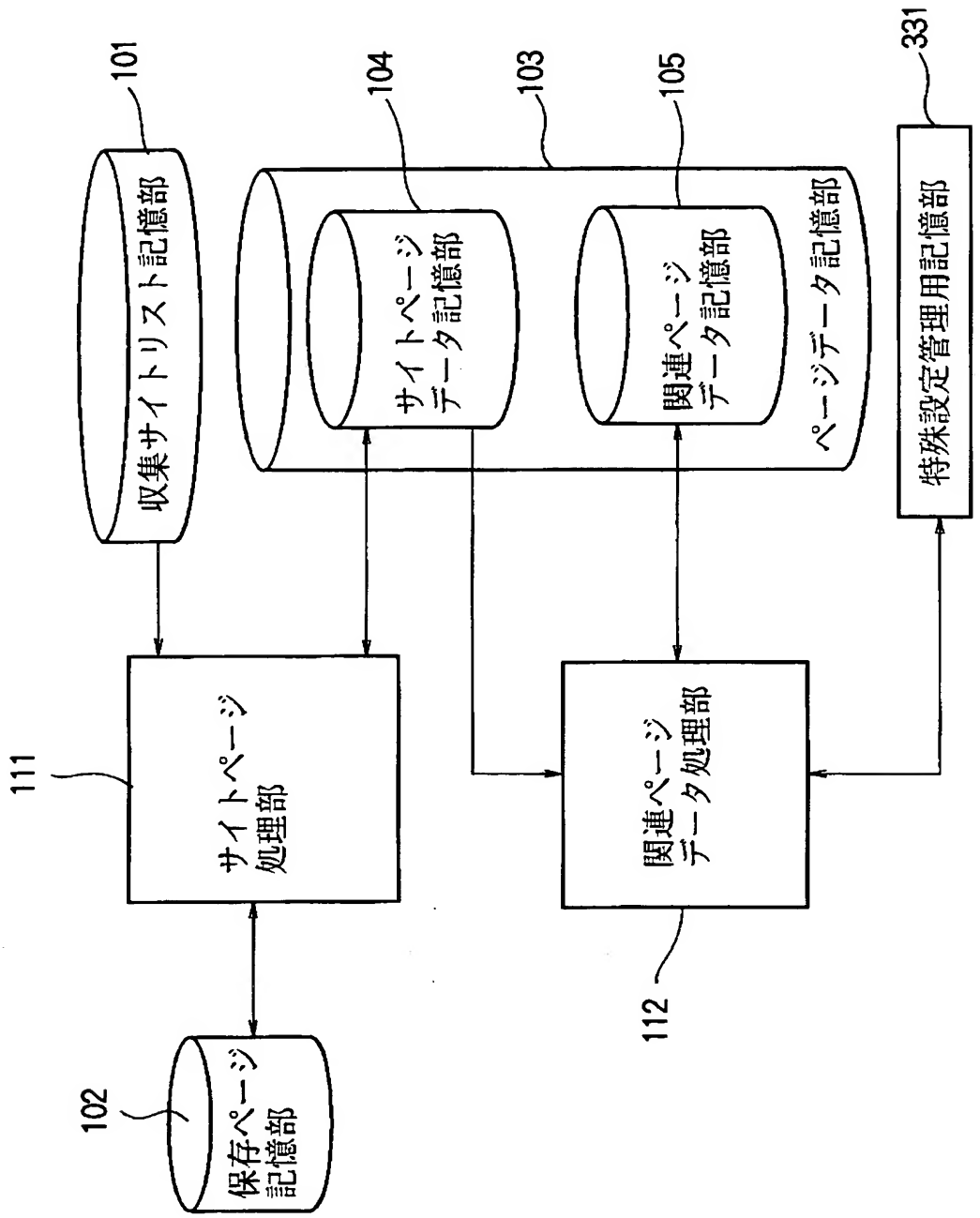
【図３１】

図31


検索処理スタート
ページ取得、保存
S101
ページID割り合て
S102
単語抽出
S103
単語ID割り合て
S104
基本ページモデル生成
S105
リンク判定
S106
SDFリンク関係情報生成
S107
CDFリンク関係情報生成
S108
SDF・CDFデータ生成
S109
ISDF・ICDFページモデル拡張処理
S110
関連度計算
S111
関連ページ一覧生成
S112
送　信
S113
エンド

【図３２】

図 32


101
収集サイトリスト記憶部
104
サイトページデータ記憶部
103
105
関連ページデータ記憶部
ページデータ記憶部
331
特殊設定管理用記憶部
111
サイトページ処理部
関連ページデータ処理部
112
102
保存ページ記憶部

【図３３】

図 33

タイトル	リンク先 URL (Key)	説明	単語	URL パターン	オーナ ID
ssss NewS	www.ssss.co.jp／News／index.html	エスエスエスエスニュース	最新情報	News	0001
ssss Present	www.ssss.co.jp／Present／index.html	エスエスエスエスプレゼント	プレゼント	Present	0002

341特殊設定用管理データ記憶部

【図３４】

図 34

オーナ ID (Key)	名前	所属	e-mail	Account	Password
0001	AAAA	HP 室	aaaa@shp.ssss.co.jp	aaaa	aaapwd
0002	SSSS	宣伝	senden@ssss.co.jp	senden	sendenpwd

342 特殊設定管理者データ記憶部

【書類名】　　要約書

【要約】

【課題】　インターネット上において、ユーザに所定のページの関連ページを精度良く提供できるようにする。

【解決手段】　サイトページ処理部１１１は、サイトに含まれるページを収集し、ページ間の親子関係を判断し、その判断結果をサイトページデータ記憶部１０４に記憶させる。関連ページデータ処理部１１２は、サイトページデータ記憶部１０４に記憶されているデータを用いて、ページ間の兄弟関係と共通親関係の少なくとも一方が考慮された重み付けが施されたページ特徴抽出の値が用いてページ間の関連度を算出する。このページ特徴抽出により、リンク関係にあるページに共通に用いられる単語が関連度算出に大きな影響を与えないように処理される。本発明は、インターネット上に設けられ、所定のページの関連ページ検索するためのサーバに適用することが可能である。

【選択図】　　図５

特願２００２－３２９４９２

出　願　人　履　歴　情　報

識別番号　　［０００００２１８５］

１．変更年月日　　１９９０年　８月３０日
［変更理由］　　新規登録
住　所　　東京都品川区北品川６丁目７番３５号
氏　名　　ソニー株式会社